



## Prosodic effects of focus and constituency in Mandarin and in English

**Wei Zhang\***, Department of Linguistics, McGill University, Montreal, Canada, [wei.zhang16@mail.mcgill.ca](mailto:wei.zhang16@mail.mcgill.ca)

**Meghan Clayards**, Department of Linguistics, McGill University, Montreal, Canada; School of Communication Sciences and Disorders, McGill University, Montreal, Canada, [meghan.clayards@mcgill.ca](mailto:meghan.clayards@mcgill.ca)

**Michael Wagner**, Department of Linguistics, McGill University, Montreal, Canada, [chael@mcgill.ca](mailto:chael@mcgill.ca)

\*Corresponding author.

---

The prosody of an utterance encodes multiple types of information simultaneously, including information status of constituents—for example, by modulations in prosodic prominence to encode focus—and information about syntactic constituent structure—by modulations of prosodic phrasing. According to many prosodic theories, however, focus and constituent structure interact with each in their effects on prominence and phrasing respectively. Focus early in an utterance is sometimes assumed to preempt the realization of tonal events later in the utterance, thus neutralizing syntactically-motivated phrasing distinctions. Other accounts assume that focus and constituent structure exert their effects on prominence and phrasing in an additive way. The current study compares English and Mandarin and investigates to what extent the correlates of focus and constituency interact with each other in shaping the prosody in production. The results show that syntax-induced phrasing distinctions are still encoded post-focally in both languages, providing new evidence for the view that different functions can be encoded orthogonally in prosody. Additionally, we found that while the two languages realize phrasing in roughly same way, they differ in their acoustic realization of focus. Mandarin relies more on F0 modulation than English, and Mandarin lexical tones interact with focus realization.

---



## 1. Introduction

Speech prosody can convey different kinds of information simultaneously. Aspects of syntactic constituent structure are encoded with prosodic phrasing, which is phonetically realized with various acoustic cues including initial strengthening and final lengthening. Semantic focus is often encoded by changes in metrical prominence, resulting in increased prominence of the focused constituent and reduced prominence for non-focal material. The type of speech act (e.g., question vs. declarative) can be encoded with the choice of intonational tune. One important research question is whether these different factors affect sentence prosody separately, or whether they interact with each other in how they shape sentence prosody. Moreover, the typological picture in the literature suggests that stress languages like English differ from tone languages like Mandarin in how focus is realized and which interactions we should expect. In this paper, we compare how focus and constituent structure affect sentence prosody in English and Mandarin.

### 1.1. Syntactic constituency and prosodic phrasing

Prosodic phrasing groups words within an utterance into larger prosodic units. One important factor that determines these chunks is syntactic constituent structure. Many studies have shown that syntactic constituency affects prosodic phrasing, and syntactic boundaries frequently coincide with prosodic boundaries (e.g., Price et al., 1994; Steedman, 1991), hence speakers and their listeners rely on the close relationship between boundary and syntax to reduce ambiguities in speech communication. Some production and perception studies point to mismatches between syntax and prosodic phrasing (e.g., Calhoun 2006; Shattuck-Hufnagel & Turk 1996; Watson & Gibson 2004), but at least some of these apparent mismatches have been argued to be based on syntactic misanalyses (Royer, 2022; Steedman 2001; Wagner, 2005).

Prosodic phrasing is often encoded by lengthening the pre-boundary syllables or segments, and/or inserting silent pauses at the phrase juncture. The final lengthening (or pre-boundary lengthening) effect is well-studied and found in many languages, although there is variation on which and how many segments preceding the boundary are lengthened (e.g., English: Klatt 1975; Turk & Shattuck-Hufnagel 2007; Wightman et al., 1992; Dutch: Cambier-Langeveld, 1997, among many others). Some studies also found initial lengthening (or post-boundary lengthening) effects on the segments immediately following the prosodic boundary, using both acoustic and articulatory evidence (in Korean: Cho & Keating, 2001; in Dutch: Cho & McQueen, 2005; in English: Pierrehumbert & Talkin, 1992). This lengthening effect at domain beginnings is thought to be a correlate of articulatory strengthening (Fougeron & Keating, 1997). In general, longer duration is able to provide more complete target realization (Lindblom, 1963). Fougeron and Keating found that the domain initial consonant and the domain final vowel are less reduced, i.e., with more extreme lingual articulations. For example, domain-final /o/ was more open than phrase medial /o/ at every boundary level for all speakers. As discussed by Cho (2015),

the lengthening effect preceding domain edges is cross-linguistically very common and may be partly physiologically motivated, but it is also controllable by speakers. The amount of both post- and pre- boundary lengthening effects varies by language (Cho, 2015; Cho & McQueen, 2005; Paschen et al., 2022).

Silent pause can be used as a cue for phrasing after relatively strong boundaries (Lin, 1999; Wang et al., 2018; Wightman et al., 1992). The Prosodic Hierarchy Theory of Prosody assumes hierarchically organized sentence structures, and proposed several categories of structures (Ladd, 2008; Selkirk, 1986), and are at the basis of the ToBI annotation (Silverman et al., 1992). One or more smaller structures can be grouped into larger structures. For example, one or more prosodic words can be clustered into an intermediate phrase, and one or more intermediate phrases can be clustered to be an intonational phrase. Stronger boundaries often show up at the edges of larger structures, and the silent pause is regarded as a cue to the stronger boundaries, such as intonational phrase boundaries or intermediate phrase boundaries (Beckman & Ayers, 1997; Li, 2002; Lin, 1999).

Apart from timing, F0 patterns around the boundaries are also possible means to signal phrasing. AM theory summarized several types of boundary tones (e.g., L%, H%) which map the F0 contour configurations onto pragmatic meanings (Ladd, 2008; Silverman et al., 1992). However, there are inconsistent views about the role that F0 plays in marking boundaries. Some studies found pre-boundary F0 lowering and/or post-boundary F0 raising effects (in Dutch: De Pijper & Sanderman, 1994; Swerts, 1997; in English: Ladd, 1988), as well as an F0 reset effect when they happen together (in German: Truckenbrodt, 2002). In other studies where factors such as lexical stress are better controlled, the contribution of F0 to mark prosodic boundaries was found to be limited (in Seoul Korean: Jeon & Nolan, 2013; in English: Lehiste, 1973; Wagner & McAuliffe, 2019). Moreover, Gollrad (2013) found that F0 contributes to intonational phrase boundaries but not to smaller phonological phrase boundaries in German, based on a series of production and perception results.

Intensity as a cue conveying prosodic phrasing within utterances has not been investigated as much as duration and F0. It is well known that there is intensity downdrift within utterances (Pierrehumbert, 1979), and a corpus study on read and spontaneous speech in Kim et al. (2006) found at least for one speaker different boundary strengths can be differentiated by intensity. It has also been found that intensity plays a role in conveying the discourse structure between utterances in Taiwan Mandarin (Tseng et al., 2009). But speakers also use intensity resets within an utterance to convey prosodic phrasing in English (Wagner & McAuliffe, 2019) and German (Poschmann & Wagner, 2016). These studies found that while intensity often decreases throughout a phrase, it can reset when a new phrase begins. The same was observed as a cue for word boundaries in a production study in Wagner (2022). It may be then that along with duration and pitch, intensity might be used as a cue for prosodic grouping across different levels of structural description.

## 1.2. Focus and its prosodic correlates

The phonological tools to encode focus prosodically have been argued to vary typologically. Kügler and Calhoun (2020), for example, identify three different ways in which languages encode focus via prosody: (i) stress-based languages, (ii) phrase-based languages, and (iii) pitch-range-based languages.<sup>1</sup>

In a stress-based language like English, the focused constituent receives metrical prominence while unfocused information is metrically reduced. The prominence of focused constituents is often assumed to encode the salience and relevance of alternatives to the constituent. These alternatives enrich the meaning of the sentence, either by changing its truth conditions (e.g., in the case of association with a focus operator like *only*), or by triggering inferences about the context that go beyond the truth conditions of the sentence itself (cf. Krifka, 2008).

Typical acoustic cues to focus in stress-based languages include increased duration, intensity, or pitch range. Modulating metrical prominence influences how pitch accents align with focused syllables resulting in higher (or lower, depending on the pitch accent type of the particular intonational contour) pitch compared to non-focal material. Post-focal material is typically compressed in pitch range (Cooper et al., 1985; Féry & Kügler, 2008). Metrical prominence also results in greater intensity and duration (as summarized in Ladd, 2008; English: Breen et al., 2010; Kochanski et al., 2005; German: Féry & Kügler, 2008; Dutch: Gussenhoven, 2004; Greek: Baltazani & Jun, 1999). In addition, focus in stress languages has been reported to be marked by hyper-articulating the distinctive segmental features, resulting in greater prosodic prominence for the focused units compared to other parts (as reviewed in Cole, 2015; Chen, 2010). See Ladd (2008) for a review and a discussion of metrical representations of prominence.

Kügler and Calhoun (2020) identify a second type of language, which encodes focus by modulations of phrasing, that is the insertion or deletion of prosodic boundaries before or after focused constituents. For example, in Korean, a boundary can be inserted before the focused constituent, while the constituents following focus are phrased together with the focus (Jeon & Nolan, 2017; Jun & Lee, 1998). Similarly, in Chichewa, a Bantu language, a boundary is inserted after the focused constituent (Kanerva, 1991). Changes in phrasing could in principle be a way of modulating metrical prominence (e.g., as argued by Büring, 2009): If occurring at the edge of a phrase itself entails metrical prominence, then phrasing changes could be a way to achieve metrical prominence and thus achieve the same goal as post-focal compression, for stress-based languages. However, Féry (2013) argues that focus-induced phrasing in fact only sometimes correlates with increased prominence on the focused constituent. Instead, Féry argues that the link between post-focal compression and phrasing changes serves to ‘align’ focus with prosodic events, and thereby help package the sentence into information-theoretically relevant domains

---

<sup>1</sup> We replaced the original term ‘register-based languages’ from Kügler and Calhoun (2020) with ‘pitch-range-based languages’ in this paper to avoid any confusion with other meanings of ‘register’ in linguistics.

that correspond to focus and topic. Either way, if focus can affect phrasing, we might expect that focus and constituent structure interact with each other in determining the prosodic phrasing of a sentence.

A third type of language, according to Kügler and Calhoun (2020), are languages that encode focus by adjustments of the register lines according to which pitch events are scaled. We will use the term pitch-range adjustment rather than register following Xu (1999). Kügler and Calhoun use Mandarin as a prime example,<sup>2</sup> since it has been reported that focus is marked by expanding pitch range on focused constituents, while it is lower and compressed on the post-focal areas (Xu 1999).<sup>3</sup> Wang et al., (2018) report that focus is signaled “mainly through pitch range adjustments, which can occur even across phrase breaks, whereas boundaries are mostly signaled by duration adjustments” (p. 24). Chen and Gussenhoven (2008), however, note that duration is also a crucial cue to focus in Mandarin. Intensity is conspicuously absent from discussions of focus realization in Mandarin, perhaps because it is often not looked at as a potential cue, and it is not listed as a cue for pitch-range-based systems in Kügler and Calhoun. This reflects a common assumption that intensity is a cue to focus only in stress-based languages. It seems then that we should expect that pitch is a more important cue to focus in pitch-range-based languages like Mandarin compared to stress-based languages, relative to other cues. On the other hand, tone languages like Mandarin have been argued to be more constrained in their use of pitch to encode focus since the lexical tonal contrasts put additional functional load on the use of pitch (see discussion in Chen & Gussenhoven). Post-focus F0 range compression was indeed found to lead to a weakened implementation of tonal targets in Chen (2010) and Chen and Gussenhoven. Chen observed that post-focus tones were more influenced by the preceding tone when realizing the tonal targets than on-focus tones, consistent with the STEM-ML model’s description of reduced prosodic strengths for the post-focal tones (Kochanski & Shih, 2003).

In sum, the literature on focus prosody suggests that languages categorically differ in their phonological means of encoding focus, ranging from modulations of metrical prominence, prosodic phrasing, or pitch range. These phonologically different means of encoding have been reported to correlate with different phonetic cues used to convey focus. Claims about typological differences in how focus is encoded need to be tested in direct comparisons between languages, and one goal of our study is to assess to what extent the typological distinction drawn between languages like Mandarin and English with respect to prosodic focus realization is justified. We are particularly interested here in how encoding focus interacts with the prosodic effects of

---

<sup>2</sup> Yoloxóchtitl Mixtec, Hindi, Akan, West Greenlandic, Georgian, Jaminjung, and Serbo-Croatian have also been argued to mark focus by modulating pitch range or pitch register (DiCanio et al., 2018; Kügler, 2020; Kügler & Calhoun, 2020).

<sup>3</sup> The direction of pitch-range effects might vary by language. Kügler and Genzel (2012) report that in Akan, focus is in fact marked by a lowered pitch.

syntactic constituency on phrasing, and whether languages differ in when syntactically-motivated phrasing distinctions are maintained post-focally.<sup>4</sup>

### 1.3. Interactions between syntactic effects and focus effects

Interactions between the effects of syntactic constituency and focus on sentence prosody are most obviously expected for languages that encode focus in a phrase-based way. If both focus and syntactic constituent structure affect the phrasing of a sentence, then we might expect to see interactions between their effects, since focus can, in principle, obscure phrasing cues to constituent structure and vice versa.

However, we might expect interactions even in stress-based and pitch-range-based languages. In stress-based languages, the correlate of focus is metrical prominence, and the acoustic cues to metrical prominence overlap with those that encode phrasing. For example, in English increases in duration can be due to prominence or phrasing. This raises the possibility that the cues to focus and constituent structure may mutually obscure each other. However, different functions may exert their effect in different locations. For example, focus prominence has greater effects on the stressed syllable of a word, while final lengthening primarily affects the last syllable. Another possibility is that the relationship between cues disambiguates their contribution. For example, Wagner and McAuliffe (2019) found that in English, when intensity and duration increase at the same time, this encodes prominence; when duration increases but intensity does not, then this provides a cue for phrasing. Whether and how cue relations serve to disambiguate the contribution of focus and constituent structure in other languages is an open question.

Depending on our assumptions about prosodic representation, we might also expect interactions due to phonological reasons. While metrical prominence and phrasing are in principle orthogonal to each other (it is possible to shift prominence while leaving phrasing intact and vice versa), some phonological theories make assumptions that suggest they are not completely orthogonal. For example, if the heads of higher-level prosodic constituents are necessarily realized by a pitch accent (as Beckman [1996] argued for Japanese), and focus, furthermore, has the effect that no pitch accents can be realized post-focally, then we would expect that post-focal phrasing distinctions should be neutralized by focus, making differences in constituent non-recoverable from the signal. Some researchers (e.g., Beckman & Ayers, 1997; Silverman et al., 1992), assume that an Intonational Phrase (ip) must be headed by a pitch accent. If focus early in an utterance preempts the realizations of pitch accents later in an utterance (see Ladd [1996] for a review of

---

<sup>4</sup> Of course, there are also languages that have been reported not to show any prosodic correlates at all (e.g., Cantonese, according to Xu, 2011). These languages may use syntactic or morphological strategies instead of prosody (Chen et al., 2016; Kalinowski, 2015).



claims for English), then we might expect that focus erases syntactically-motivated phrasing in the post-focal domain that relies on ip-phrasing.

However, some phonological accounts of metrical prominence and phrasing assume that they are, in principle, orthogonal, and hence we do not necessarily expect post-focal neutralization of syntactically-motivated phrasing distinctions (see Wagner & McAuliffe, 2019, for discussion and a concrete example). Similarly, prosodic models that were created to explicitly predict and synthesize speech prosody from texts (SFC model, Bailly & Holm, 2005; CR/Fujisaki model, Fujisaki, 1983; PENTA model, Xu, 2005; INTSINT model, Hirst & Espesser, 1993; STEM model, Kochanski & Shih, 2003; TILT model, Taylor 2000; linear alignment model, Van Santen & Möbius, 2000) often assume that functions like focus and constituent structure exert their separate prosodic effects in an additive way.

There is indeed evidence that focus does not erase post-focal phrasing distinctions, at least in languages that have been characterized as stress-based, such as English and German. Pierrehumbert (1980, p. 223) already reported that pitch accents in English can be realized after the nuclear accents, albeit with a compressed pitch range, thus opening the door that pitch cues to phrasing distinctions in the post-focal domain may be maintained after all. Norcliffe and Jaeger (2005) found that at least durational cues to post-focal phrasing reflecting syntactic constituency were maintained. Similarly, Kügler and Féry (2017) investigated the post-focal downstep effect in German and found that in the post-focal domain, there was an extremely compressed range, but phrasing distinctions were not neutralized. Wagner and McAuliffe (2019) used coordinate structures of names and varied focus position and phrasing, for example, '(Megan and [Dillon]<sub>FOCUS</sub>) or Morgan would help' and '[Megan]<sub>FOCUS</sub> and (Dillon or Morgan) would help'. Results showed that post-focal phrasing remains intact in English. While focus affects the phonetic cues used to convey phrasing (for example, via pitch compression), it does not altogether erase information about syntactically-motivated phrasing distinctions. Related results for English are reported in Wu (2021).

For languages such as Mandarin that have been classified as pitch-range-based, it is less clear how exactly we might expect focus and syntactically-motivated phrasing to interact. If intensity is indeed not used to mark focus in languages like Mandarin, or at least used less frequently, and if pitch range reduction is more limited in tone languages (Chen & Gussenhoven, 2008), this could influence the manner and extent to which post-focal phrasing is affected, compared to stress-based languages. For example, if Wang et al. (2018) are correct that duration is mostly used to encode phrasing in Mandarin and less so for focus, then we might expect that it is easier to preserve post-focal phrasing cues in Mandarin compared to English, where both cues have been reported to be important for encoding both focus and phrasing. On the other hand, Chen and Gussenhoven have argued that duration is important for focus as well as phrasing, which could limit the ability to code both factors in parallel.

Most studies that inform the current typological picture of the prosodic realization of focus are based on studies on individual languages, while direct comparisons between languages remain scant (see Calhoun et al., 2021, and Yan & Calhoun, 2020, for two examples). When languages are directly compared to test the validity of typological claims, the results are sometimes surprising. Kügler and Calhoun (2020) report that phrase-based focus realization correlates with lack of lexical stress, and that those languages lack the ability to shift prominence elsewhere, citing French as an example. Vander Kloek et al. (2018), however, directly compared prosodic focus marking in English (argued to be a stressed-based language) and French (argued to be a phrasing-based language), and found that prosodic focus marking is phonetically remarkably similar in the two languages—focused constituents are boosted in duration, pitch, and intensity, and post-focal material is reduced. The evidence calls into question the claim that in French it is not possible to phonologically deaccent material within phrases (Féry, 2014). Where French clearly does differ from English, however, is when focus prosody is used: Prosodic focus was only reliably encoded for corrective focus, confirming earlier observations in Ladd (1990) and Cruttenden (1994), while it is rarely used to mark parallelism within a sentence. In other words, English and French mostly differ when prosody is used, and less in how it is phonologically or phonetically implemented. Similarly, Hamlaoui et al. (2019) looked at Polish and Czech, two other languages without lexical word stress (and thus similar to French), and found focus marking to be very similar to Germanic, and less restricted than in Romance languages including French. This suggests that there is no correlation between lack of contrastive lexical stress at the word level and focus realization at the phrasal level, contrary to the typological assumptions often made, including those in the review by Kügler and Calhoun.

There have not been many studies that directly compare stress-based and pitch-range-based languages (see Wang et al. [2019] for one example of difference in use of durational cues to prosodic boundaries in Mandarin and American English). This study aims to directly compare a stress-based language, American English, and a pitch-range-based language, Mandarin, to further test the validity of current typological assumptions about language types, how they encode focus and constituent structure and how their interaction with each other affects sentence prosody.

#### **1.4. Interaction between lexicon and speech prosody**

When comparing languages at the level of sentence prosody, it is important to take into account lexical factors that distinguish them. In American English, domain-initial strengthening impacts unstressed initial syllables to a greater extent than stressed initial syllables (Kim et al., 2018), exhibiting the interaction between word-specific phonological content and the phonetic implementation of phrasing. The effect of word-specific phonological content on the phonetic implementation of phrasing seems to be language-specific. Cambier-Langeveld (1999) compared the interaction between final lengthening and accentual lengthening in Dutch and English, and



found that there was significant accentual lengthening only in non-final positions in Dutch. However, in English, the accentual lengthening is consistent across all types of positions.

In languages without lexical stress, accentual lengthening showed differences from the patterns in languages with lexical stress (Seo et al., 2019; Tsai, Jang et al., 2020; Tsai & Katsika, 2020). For instance, Seo et al. found that in Japanese, when the initial syllable of a disyllabic word was pitch accented, the pre-boundary lengthening effect on the final syllable was suppressed, in contrast to in English or Greek where the final syllable is still lengthened (e.g., Katsika, 2016; Turk & Shattuck-Hufnagel, 2007).

In Mandarin, we might expect effects of lexical tone on prominence and phrasing, which could modulate the effects of focus and constituent structure, respectively. An important difference between English and Mandarin is that English has word stress whereas Mandarin has lexical tones. Mandarin has four full tones which are distinguished by F0 contours, high-flat (T1), low-rising (T2), low-dipping (T3), high-falling (T4), and a neutral tone (T0), which correlates to weak syllables (Chao, 1965). English and Mandarin may exhibit differences in their marking of focus on the F0 dimension, partly because lexical tone distinction may need to be conveyed, also post-focally in Mandarin, but not in English (Xu, 1999).

### **1.5. The current study**

In this study, we aim to compare the realization of focus and constituency in Mandarin and English. Our main interest in this paper is on the following three research questions. The first is whether focus neutralizes phrasing distinctions motivated by constituent structure in the post-focal domain. Previous findings in English and German (Norcliffe & Jaeger, 2005; Wagner & McAuliffe, 2019; Wu, 2021) suggest that post-focal phrasing distinctions are maintained, and this study will help assess whether the same is true in Mandarin.

The second research question regards the role of pitch and intensity in marking focus and constituency. English, just like Mandarin, uses pitch range expansion and post-focal pitch range reduction to encode focus, but Mandarin has been assumed to rely more on pitch compared to other cues, such as duration and intensity. However, Mandarin, even if viewed as a pitch-range-based language, has been reported to employ intensity as a cue to focus, which is more typically associated with stress-based languages (Shih 1988) and duration (Chen & Gussenhoven, 2008). Chen and Gussenhoven, in fact, argued that post-focal pitch range reduction in Mandarin is more limited so that lexical tone distinctions can be maintained. A direct comparison between the languages will help clarify how the cues are used and weighted in relation to each other. In English, intensity is also a crucial cue to phrasing (Wagner & McAuliffe 2019; Wagner 2012), but is the same true in Mandarin? Wang et al. (2018) hypothesized a role for intensity in marking constituency in Mandarin. Intensity has been reported to play a role in marking discourse

structure in Mandarin (Tseng et al., 2009), but is it also involved in encoding grouping into phrases and words within sentences, as in English?

A third research question is how the choice of lexical tone affects the realization of prosodic cues to focus and constituency. Mandarin tones contain differing targets. T1, for example, contains only the high target, whereas T2 and T4 contain both high and low targets. Focus marking is observed to be dependent on the tonal targets (Wang et al., 2020; Xu, 1999), but how differing tonal targets influence the realization of constituency remains less clear. This study will explore the tonal effect on the prosodic encoding of focus and constituency. To make parallel comparisons between Mandarin and English, the effect of lexical stress patterns in English will also be investigated.

## 2. Method

To investigate the interaction of different factors affecting sentence prosody in Mandarin and English, we conducted production experiments in which we manipulated focus and constituency. All data, analysis scripts, experimental materials and preregistration are available at <https://osf.io/2wnjm/>.

### 2.1. Materials

We used coordinate structures as the production materials, following Wagner and McAuliffe (2019). Three names (represented by A, B and C below) were connected by the two conjunctions ‘or’ and ‘and’:

(1) A or B and C

Our constituency manipulation thus had two conditions: either left branching as in (2), where the first two names connected by ‘or’ are in the same constituent; or right branching as in (3), where the last two names connected by ‘and’ are in the same constituent.

(2) [A or B] and C

(3) A or [B and C]

We manipulated four conditions for focus: either one of the three names (A, B or C) was focused, or there was wide focus (no specific name is focused). The target productions were designed to answer a question about who did something in a dialogue, and the coordinate structure was embedded in the answer. The answer sentence was elicited by context information (text and audio, as well as a figure) which suggested the constituency and focus conditions. The focus was encoded to correct a name mentioned in the question, and the constituency was encoded to show the person’s affiliation. Wagner and McAuliffe (2019) used a reading task, and encoded

constituency within the coordinate structure using commas. Commas, however, might also be interpreted by readers as directly encoding prosodic phrasing, in addition to constituency. In the study here, we aimed for a more naturalistic task, where speakers needed to assemble their response based on scenario and visualization aimed at eliciting the intended constituency and focus conditions without punctuation (see also **Table 1**). The names were all bi-syllabic but differed in the lexical patterns in Mandarin (varying in lexical tones) and in English (varying in lexical stress), as described in sections 2.1.1 and 2.1.2.

Constituency and focus manipulations were crossed for 8 conditions. Each condition occurred with 2 levels of lexical tone of the name sets (in the Mandarin experiment), or 2 levels of stress patterns of the name sets (in the English experiment), and each condition had four repetitions. There were 64 trials in each experiment (2 levels of Constituency, 4 levels of Focus, 2 levels of Lexical tone/Stress pattern, \*4 repetitions). For variety, these 8 focus-constituency conditions occurred in 8 different scenarios, and were blocked into sets of 8 trials with a single scenario (e.g., going to a game). Each block of 8 trials included all combinations of constituency and focus. Half of the trials used names with one tone/stress type and the other half used the other tone/stress type (see **Table 1** for an example).

### 2.1.1. Mandarin materials

The names used in the Mandarin experiment were composed of a monosyllabic Chinese last name, followed by a monosyllabic Chinese honorific ‘Ge1’, which means elder brother. (‘Ge1’ is the PINYIN form. This corresponds to /kʰ/ with T1.) Together, these two syllables form a disyllabic Chinese name. The last names are either with T1, the high-flat tone, or T4, the high-falling tone. They were selected because they contain the highest tonal targets in Mandarin tonal space, and high targets are thought to show more similarity to English focus realization, compared to low targets (Zhang et al., 2008). In addition, the other two tones were avoided because of tone sandhi rules, which would complicate the interpretation.

A scenario containing eight conditions of the Mandarin experiment is listed in **Table 1** for illustration. For example, the context for the left branching and A focus condition was “*Who went to clean up the classroom? Xiao3Ming2 yesterday said Su1Ge1 or Cui1Ge1 in Class A went to clean up the classroom, and took along Fang1Ge1 in Class B.*” Here, “*Su1Ge1 or Cui1Ge1 in Class A ...*” indicates the Left branching condition. The bracketing was also expressed by the distances between the characters in the figures (the first two characters were closer to each other than to the third one). The (corrective) focus condition was expressed in the figure as well, by a red X. This example was initial focus (A was the focus) indicated by a red X on the Su1Ge1 character image, and the correct character—Xiao1Ge1—was shown below. The target sentence (spoken by the participants) for this condition was “*Not Su1Ge1 or Cui1Ge1 and Fang1Ge1 who went, but Xiao1Ge1 or Cui1Ge1 and Fang1Ge1 who went.*” The underlined portion in of the answer was extracted and analyzed.

Focus	Left Branching Constituency			Right Branching Constituency		
	Context	Figure	Target	Context	Figure	Target
Wide	Who cleaned up the classroom? I heard from Xiao3Ming2 that Wang1Ge1 did it.		Not Wang1Ge1 who went. It was <u>Jiang1Ge1 or Jin1Ge1 and Qiu1Ge1</u> who went.	Who went to clean up the classroom? Xiao3Ming2 yesterday said Song4Ge1 went to clean up the classroom.		Not Song4Ge1 who went. It was <u>Dou4Ge1 or Jin4Ge1 and Cai4Ge1</u> who went.
A	Who went to clean up the classroom? Xiao3Ming2 yesterday said Su1Ge1 or Cui1Ge1 in Class A went to clean up the classroom, and took along Fang1Ge1 in Class B.		Not Su1Ge1 or Cui1Ge1 and Fang1Ge1 who went, but <u>Xiao1Ge1 or Cui1Ge1 and Fang1Ge1</u> who went.	Who went to clean up the classroom? Xiao3Ming2 yesterday said either Meng4Ge1 in Class A went to clean up the classroom, or Dai4Ge1 and Luo4Ge1 in Class B together went to clean up the classroom.		Not Meng4Ge1 or Dai4Ge1 and Luo4Ge1 who went. It was <u>Xia4Ge1 or Dai4Ge1 and Luo4Ge1</u> who went.
B	Who went to clean up the classroom? Xiao3Ming2 yesterday said Zhao4Ge1 or Du1Ge1 in Class A went to clean up the classroom, and took along Fan4Ge1 in Class B.		Not Zhao4Ge1 or Du1Ge1 and Fan4Ge1 who went. It was <u>Zhao4Ge1 or Huo4Ge1 and Fan4Ge1</u> who went.	Who went to clean up the classroom? Xiao3Ming2 yesterday said either Ding1Ge1 in Class A went to clean up the classroom, or Sun1Ge1 and Zhou1Ge1 in Class B together went to clean up the classroom.		Not Ding1Ge1 or Sun1Ge1 and Zhou1Ge1 who went. It was <u>Ding1Ge1 or Zhang1Ge1 and Zhou1Ge1</u> who went.
C	Who went to clean up the classroom? Xiao3Ming2 yesterday said Fei4Ge1 or Duan4Ge1 in Class A went to clean up the classroom, and took along Zheng4Ge1 in Class B.		Not Fei4Ge1 or Duan4Ge1 and Zheng4Ge1 who went. It was <u>Fei4Ge1 or Duan4Ge1 and Jing4Ge1</u> who went.	Who went to clean up the classroom? Xiao3Ming2 yesterday said either Xin1Ge1 in Class A went to clean up the classroom, or Pan1Ge1 and Zhu1Ge1 in Class B together went to clean up the classroom.		Not Xin1Ge1 or Pan1Ge1 and Zhu1Ge1 who went. It was <u>Xin1Ge1 or Pan1Ge1 and Gao1Ge1</u> who went.

**Table 1:** Example of the context, paired figure and target sentence for eight conditions within one scenario (cleaning the classroom). The underlined part of the target sentence was analyzed.

The eight sets of Mandarin last names and eight sets of English names are listed in Table S1 in the supplementary materials. These names were matched to avoid tongue twister-like productions. Although consonants and vowels could not be strictly controlled across A, B and C, each name position, i.e., each column in Table S1, included high, mid and low vowels. This distribution is used to reduce the intrinsic F0 effects in the vowels (Whalen & Levitt, 1994). The full set of context and target sentences are available at <https://osf.io/2wnjm/>. All Mandarin context sentences were pre-recorded by a native speaker of Mandarin (the first author).

### **2.1.2. English materials**

The names used in the English experiment were bi-syllabic. Since the stress pattern may affect the marking of focus and constituency, we controlled the stress patterns of the English names: Four sets of names were initial-stressed (e.g., Lauren) and four sets of names were final-stressed (e.g., Nichole). The scenarios were translated into English from the Mandarin ones, with English target names substituted for the Mandarin names.

In summary, there were 64 trials in each experiment. In the Mandarin experiment there were 2 constituency levels \* 4 focus levels \* 2 tone types \* 4 repetitions. In the parallel English experiment there were 2 constituency levels \* 4 focus levels \* 2 stress patterns \* 4 repetitions. The full set of context and target sentences are available at <https://osf.io/2wnjm/>. English context sentences were pre-recorded by a native speaker of English (the second author).

## **2.2. Participants**

### **2.2.1. Mandarin participants**

The Mandarin experiment was carried out in Nanjing, China. Twenty native Mandarin speakers were recruited by social media advertisements directed to college students. Participants were unfamiliar with linguistic concepts, and none of them reported any hearing, reading, or speech issues. Participants were invited to the recording room to do the experiment.

Participants were compensated for their time. After exclusion (one was excluded because the experimenter used different instructions; two were excluded because of extraneous noise resulting from equipment issues), seventeen participants' data were used in the analysis (nine female, eight male). Eight participants reported growing up in the northern part of mainland China, whereas nine were from the southern part.

### **2.2.2. English Participants**

Due to the COVID-19 pandemic, the English experiment was carried out online. Twenty-three native English speakers were recruited through the online platform Prolific. Two were excluded because their production didn't follow the target sentences. Another four were excluded because

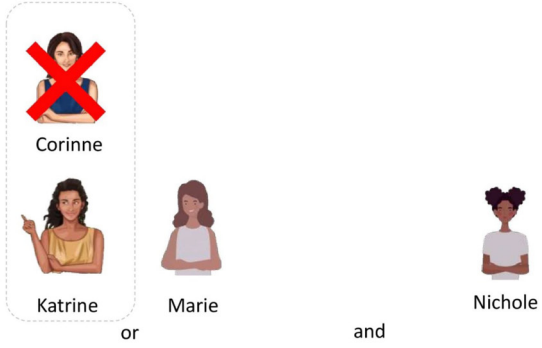
they reported growing up in areas (Nigeria, Ireland, Jamaica, and Philippines) other than North America. Seventeen participants' data were used for analysis (nine female, eight male).

### 2.3. Procedure

The experiment was conducted using the prosodylab experimenter (Wagner, 2021), a set of javascripts building on jspsych (De Leeuw, 2015).

One each trial, participants were first shown a screen with the main information in **Table 1**: the context (top), a cartoon figure of the scenario (middle), and the target sentence (bottom). An example is given in **Figure 1**. This allowed the participants to familiarize themselves with the scenario. When they were clear about the scenario and ready to speak, they clicked a button at the bottom of the screen. Participants then heard the pre-recorded audio of the context question and were asked to reply with the target sentence—using their understanding of the scenario—as if they were talking to the person in the audio. Their responses were recorded automatically. All the information (contexts, figures, targets) remained on screen during the dialogue recording. The dialogue format was designed to elicit natural responses from the participants. After each recording, the recorded audio was immediately played back, giving participants the option to redo the recording if they made a mistake. There were three practice trials for the participants to familiarize themselves with the experiment format. After the practice, there were 64 experimental trials.

Who went to the game? John said yesterday that Corinne or Marie in Class A went to the game, and took along Nichole in Class B



Corinne

Katrine or Marie and Nichole

**It's not Corinne or Marie and Nichole who went. It was Katrine or Marie and Nichole who went.**

*Listen, then say your response!*

**Figure 1:** An example of the second screen in an English trial, including the context, visual prompt, target sentence and a brief instruction.



In the Mandarin experiment, stimuli were presented on the computer screen and through AKG K271 MKII headphones. Audio was recorded using a cardioid condenser microphone Neumann U87Ai and the audio interface RME Fireface 800. The experimenter (the first author) monitored the experiment using another set of headphones and another synchronized screen outside the soundproof room, in case of any problems during the experiment. In the online English experiment, however, participants were not monitored and used their own headsets with a microphone.

## 2.4. Acoustic measures and statistical models

The audio files were aligned with the Montreal Forced Aligner (McAuliffe et al., 2017). A set of acoustic features were extracted by a Praat script.<sup>5</sup> In the script, F0 extraction followed a two-step method to reduce F0 tracking errors. In the first step, the default settings were used and the ceiling and floor F0 were calculated from the resulting measurements. In the second step, the default settings were replaced with these ceiling and floor values to ensure a talker-specific F0 extraction. For each syllable in both languages, we extracted the following acoustic correlates: syllable duration, maximum F0, and mean intensity. The maximum F0, instead of mean or minimum F0, was selected because the two Mandarin tones we used have asymmetric tonal targets, and they both have the high target in the tonal onset. Hence, by maximum F0, one can investigate the language-specific marking pattern instead of the tone-specific pattern in Mandarin. The use of mean intensity followed Wagner and McAuliffe (2019). There were eight syllables in the coordinate structure so that there were 24 correlates for each utterance.

Random forest and linear mixed-effect models (LMM) were used for statistical analysis in this study. The statistical analyses were run on the R platform (R Core Team, 2013) via *party* and *lme4* packages (Bates et al., 2015; Hothorn et al., 2010).

### 2.4.1. Random forest models

To investigate which features are best at encoding focus and constituency, random forest models were fit to classify the utterances, according to either focus or constituency, using all acoustic correlates. These models provided a ranking of the acoustic correlates based on their relative contribution to the classification of either the focus or constituency conditions. Random forest models are found to be relatively robust when there is collinearity among the predictors and can process a larger number of predictors, compared to other statistical methods (Strobl et al., 2009; cited after Wagner & McAuliffe, 2019). Four random forest models were fit in this study. For each language, one random forest model was fit for focus classification and the other for constituency classification. In the random forest models of focus classification, eighteen features (F0, duration and intensity for the three bi-syllabic names) were used. In the random forest

---

<sup>5</sup> Retrieved from <https://github.com/prosodylab/prosodylab.praatscripts>.

models of constituency classification, all 24 features (F0, duration and intensity for all eight syllables) were used instead, since the realization of the coordinators (*and* and *or*) in the target sentence were also found to mark constituency in the empirical plots.

If constituency is encoded by phrasing, and if phrasing distinctions are neutralized in the post-focal domain, then we expect that constituency will not be successfully conveyed in the post-focal domain. To examine this, the constituency accuracy for each focus condition was analyzed. If constituency is not recoverable post-focally, the constituency accuracy in the condition of initial focus should be lower than in the other conditions. On the other hand, if initial focus does not actually neutralize post-focal phrasing-correlates of constituency, then there should be no such pattern.

#### 2.4.2. Linear mixed-effect models

Linear mixed-effect models (LMM) were used to test the effects of focus, constituency, and their interaction on each cue. For each language, we fit models for the 3 acoustic cues and the 2 syllables in the names separately. Hence there were 6 models (3 cues: F0/intensity/duration \* 2 syllables: initial/final syllable) for each language. We also fit models to a mix of English and Mandarin data, which will be explained in more detail in Section 3.3. One of the three acoustic cues was the dependent variable in each model, and the fixed predictors encoding the focus and constituency were as follows.

The first three predictors reflected our hypotheses about the effects of focus. Each syllable had one of four levels of focus (narrow focus, wide focus, no focus before a narrow focused syllable and no focus after a narrow focused syllable), allowing us to ask three questions. Our first contrast (others.vs.noFocus) tests whether the syllables that had either narrow (A, B or C) or wide focus ('others') were different from syllables when focus was on some other constituent ('noFocus'). The next contrast separates wide focus from narrow focus (wide.vs.focus). These two contrasts are orthogonal to each other and are a partial Helmert coding of the four levels.

- others.vs.noFocus: The value is 1 if the syllable is not the focus (but the utterance is not wide focus). The value is -1 if the syllable is the focus or its utterance is wide focus.
- wide.vs.focus: The value is 1 if the syllable (the name) is the focus. The value is 0 if the syllable is not the focus (but the utterance is not wide focus). The value is -1 if its utterance is wide focus.

Our third contrast allows us to ask whether out of focus syllables that occur before the focused constituent are different from those that occur after the focused constituent. This comparison is crucial to establish whether there is post-focal compression (and not just some more general reduction that affects constituents before and after the focus). This last contrast is not fully orthogonal to the previous two, but necessary given our research questions.

- *preFocus.vs.postFocus*: The value is 1 if the syllable is preceded by a focused name. The value is -1 if the syllable is followed by a focused name. The value is 0 if the syllable is the focus or its utterance is wide focus.

Our next set of contrasts concerns where the syllable falls with respect to the intended constituent structure. We coded the syllables according to whether a constituent boundary was intended to follow them or not. There are three levels (no boundary, constituent boundary, phrase boundary). The first contrast (*final.vs.others*) compared name C to those not at a phrase edge. The second one (*noBoundary.vs.boundaryInternal*) tests whether phrase internal constituent boundaries had an effect on syllables just before them. These two codings are orthogonal to each other.

- *final.vs.others*: The value is 1 when the syllable is not in name C and the value is -1 if the syllable is in name C.
- *noBoundary.vs.boundaryInternal*: The value is 1 if there is a sentence internal boundary (induced by Left or Right branching) to the right of the name the syllable belongs to (e.g., syllables in name B when Left branching). The value is -1 when there is no sentence internal boundary (induced by Left or Right branching) to the right of the name the syllable belongs to. The value is 0 when the boundary to the right of the name is the boundary of the coordinated structure, i.e., when the syllable is in the name C.

We also included several other variables:

- *position*: a numeric variable encoding the real ordinal number of syllable in the utterance. This was added to account for the declination effect over the utterance.
- *T4.vs.T1* (only used in Mandarin models): contrasting the tonal type of the last name the syllable belongs to. The value is 1 if the Chinese last name is T1 and is -1 if it is T4.
- *FN\_S.vs.IN\_S* (only used in English models): contrasting the stress pattern of the name the syllable belongs to. The value is 1 if the English name is initial-stressed and is -1 if it's final-stressed.

Each model also included three two-way interaction terms. Most importantly, the effect *noBoundary.vs.boundaryInternal\*preFocus.vs.postFocus* was added to examine focus-constituency interactions in the productions. If the syntactically-motivated post-focal prosodic phrasing is maintained, we expect the interaction to be non-significant in the models of all three cues for both syllables and both languages. For Mandarin models, the effects *T4.vs.T1\*noBoundary.vs.boundaryInternal* and *T4.vs.T1\*others.vs.noFocus* were added. For English models, the effects *FN\_S.vs.IN\_S\*noBoundary.vs.boundaryInternal* and *FN\_S.vs.IN\_S\*others.vs.noFocus* were added. These interactions tested whether the tonal target or the stress pattern interacted with the marking of constituent structure or focus.

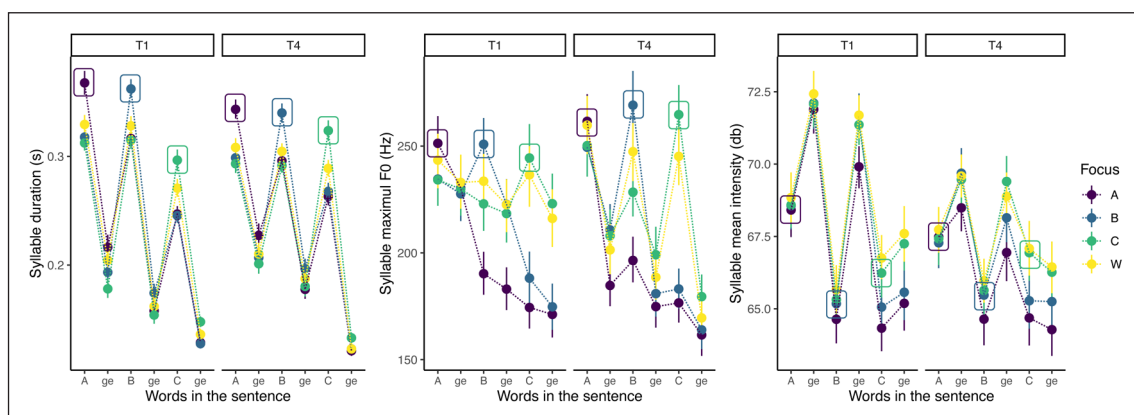
The predictors were named such that the first level mentioned is coded with a negative value (e.g., final vs. others, final is  $-1$ ) and the second level has a positive value (other is  $+1$ ). Therefore, a positive effect estimate for the factor indicates an increase in dependent variable for the second level relative to the first.

Each of the fixed effects was standardized by centering and dividing by two standard deviations. In each model, by-participant random effects were used. We first added the random intercept and the random slopes of the three focus-related fixed effects, two constituency-related fixed effects and the focus-constituency interaction effect. Then for each model, we used the *rand()* function in *lmerTest* (Kuznetsova et al., 2017) package to select the random slopes that significantly improved the model, and only used those in the final models.

### 3. Results

#### 3.1.1. Focus realization (Mandarin)

**Figure 2** shows the focus effect on the 3 acoustic cues—syllable duration, maximum F0 (max F0 hereafter), mean intensity—of the three target names, grouped by tone type of the last names used in the utterance.



**Figure 2:** Syllable duration, max F0 and intensity of the target names, pooled by focus condition (indicated by colors) and tone types of the last names (split by panels). Rectangles label the first syllable (i.e., the last name) of the focused name for each focus condition. In the legend, A, B, C, and W represent focus on A Ge, B Ge, C Ge, and wide focus, respectively.

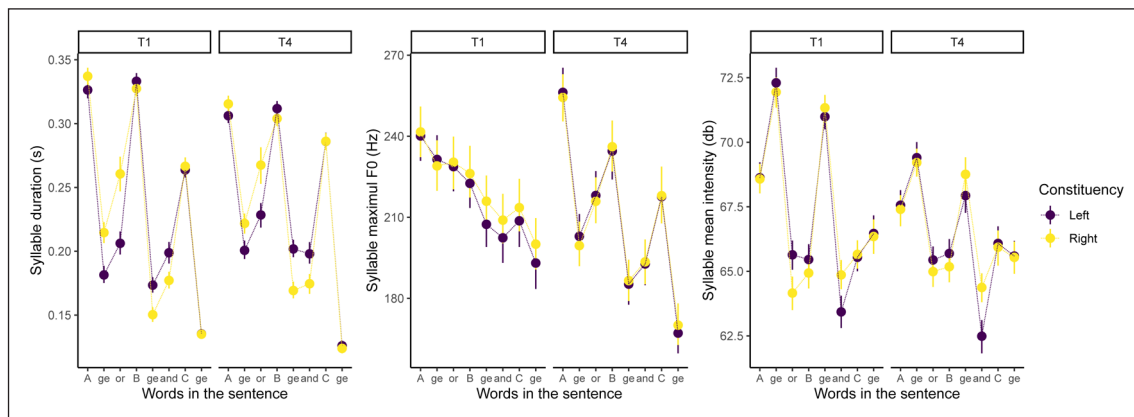
From **Figure 2**, it seems that both syllable duration and max F0 are informative cues for focus. The focused words are longer and have higher max F0. The duration of the last names is, on average, 34 ms longer than their duration in the wide focus condition, averaging over constituency and tone type. Duration of the ‘ge’s is 17 ms higher under focus than they are in the wide focus condition, averaging over constituency and tone type. Max F0 marks the focus mainly through the first syllable in the name. For the last names, max F0 is 13 Hz higher under

focus than in the wide focus condition. For ‘ge’s, however, Max F0 is 2 Hz lower than wide focus condition. Additionally, in the post-focal domain, there is a notable F0 decrease. For example, when A is on focus, the Max F0 is 63 Hz lower on B than on A, averaging over constituency and tone type.

T1 and T4 show different patterns in all three acoustic cues. Overall, T1 has larger ranges of variation on duration and intensity than T4 over the utterance. T4 has larger ranges of variation on max F0 than T1, even though they both have a high target in the citation form.

### 3.1.2. Effects of constituency (Mandarin)

Figure 3 shows the constituency effect on the 3 acoustic cues (syllable duration, max F0, mean intensity) of the coordinate structure, grouped by tone type of the last names used in the utterance. For both tone types, duration marks the constituency conditions. Our hypothesis is that constituency is encoded prosodically via phrasing. For example, the Left branching condition should induce a boundary to the right of *B\_ge*, and this is consistent with *B\_ge* having a longer duration by 26 ms in the Left branching than when it is in the Right branching condition, and *and* having a longer duration by 11 ms than when it is in the Right branching condition, averaging over focus and tone types.



**Figure 3:** Syllable duration, max F0 and intensity for Left and Right branching conditions, grouped by tone types of the last names. Panels are the same as in Figure 2.

The Max F0 is 4 Hz lower for *B\_ge*, and 5 Hz lower for *and* in the Left branching condition than in the Right branching condition for T1 names, averaging over focus and tone types. For T4 names, the trend is very weak so it seems that Max F0 doesn't contribute to phrasing marking for T4 names.

Intensity seems to mark constituency-induced phrasing. For example, the *B\_ge* and *and* are of lower intensity in the Left branching condition than in the Right branching condition. Specifically, the difference is 1.2 dB for *B\_ge* and 1.7 dB for *and*.

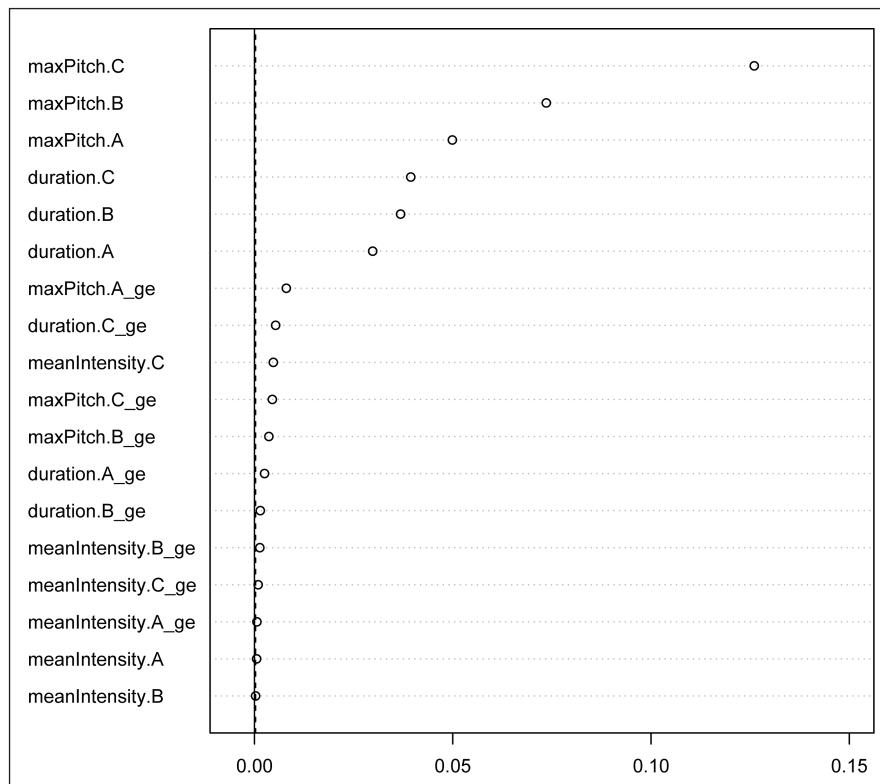
Overall, as observed in focus realizations, T1 and T4 show differences in the three cues. T1 has larger variations in syllable duration and mean intensity and smaller variations in max F0 than T4.

### 3.1.3. Random forest results (Mandarin)

To investigate the relative contribution of each acoustic feature at each position when marking the focus and constituency in Mandarin, we fit one random forest model for focus and one for constituency.

#### 3.1.3.1. Random forest model of focus classification (Mandarin)

The features' relative contributions to focus classification are shown in **Figure 4**. The solid black line marks the zero values of the x-axis. In principle, points on the black line show zero-effect on the classification. The dashed line shows the absolute value of the lowest negative predictor. Any points to the right of the dashed line can be regarded as contributing to the model accuracy, while predictors to the left possibly do not (Strobl et al., 2009; cited after Wagner & McAuliffe 2019). The farther the point is from the dashed line, the more this predictor contributes to the classification accuracy.



**Figure 4:** Relative contribution of the features in the random forest model of focus classification (Mandarin).



**Figure 4** shows that the most important cues for focus are max F0 and the duration of the last names. The next contributive cues are the max F0 and duration of the *ge* syllables. The overall classification accuracy (out-of-bag prediction) of the random forest was 65.9%, and the confusion matrix is shown in **Table 2**. The initial focus condition has the highest accuracy (78.9%), whereas the wide focus condition has the lowest accuracy (46.6%). All conditions are classified correctly above chance (25%). The wide focus condition was most often confused with the ‘C Ge1’ focus condition, suggesting that the phrase final words of the two conditions were realized in similar ways.

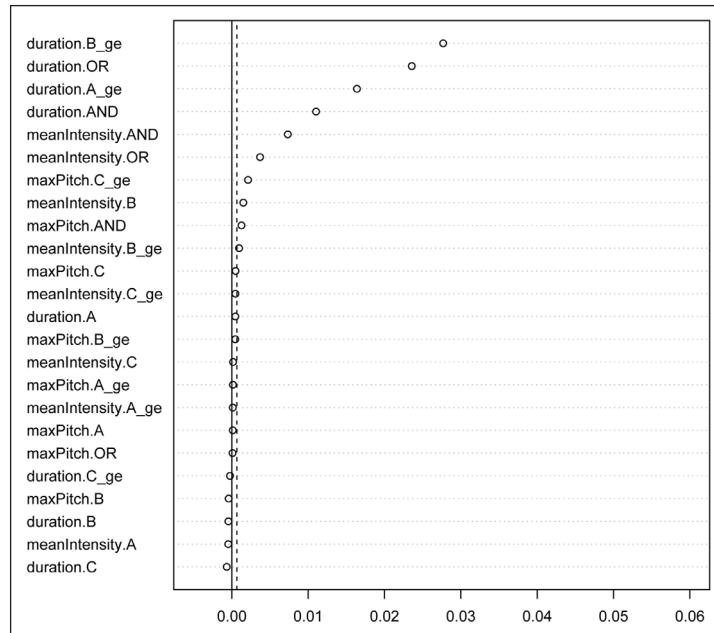
	A	B	C	W	Accuracy
A	210	24	12	20	79%
B	26	189	18	31	72%
C	12	12	180	67	66%
W	26	40	77	125	47%

**Table 2:** Confusion matrix of the random forest model for focus classification (Mandarin). Row labels indicate the intended focus condition. Column labels indicate the classified focus condition. Chance level is 25%.

### 3.1.3.2 Random forest model of constituency classification (Mandarin)

**Figure 5** shows the relative contribution of features in the random forest model for the constituency classification. The overall classification accuracy (out-of-bag prediction) was 63%. **Figure 5** shows that duration is the most informative cue for constituency classification. Durations of the syllables preceding the internal boundaries (duration.B<sub>ge</sub> for Left branching, and duration.OR for Right branching) contributed the most to the classification. Features of duration.A<sub>ge</sub>, and duration.AND were also important contributors. Other than duration, intensity of the post-boundary syllables also contributed to the classification. The result is consistent with Wagner and McAuliffe (2019), in that constituency is marked though both duration and intensity, as expected if constituency is encoded through prosodic phrasing and if duration and intensity are each reliable cue to prosodic phrasing. F0 seemed to contribute to encoding constituency as well, but the size of effect was very small.

To examine whether constituency is recoverable in post-focal position in Mandarin, we combined the classification results of focus and constituency together and analyzed the constituency accuracy for each focus condition. If, as predicted in Section 2.4.1, prosodic cues to constituency are neutralized post-focally because phrasing distinctions are lost, then the classification accuracy for constituency in the initial focus case should be lower than in the other conditions, unless initial focus is not actually realized. So, for those trials where initial focus is accurately classified in the random forest model, the constituency accuracy should be low, whereas when the initial focus is not accurately classified, the constituency accuracy should be higher.



**Figure 5:** Relative contribution of the features in the random forest model for classifying constituency (Mandarin).

The constituency accuracy for each focus classification state is shown in **Table 3**. The highest constituency classification accuracy is in the wide focus condition. The constituency accuracy in the initial focus condition is not the lowest across all focus types, rather, it actually seems to be higher than B-focus and C-focus conditions. Crucially, in the initial focus condition, constituency classification accuracy is *higher* (0.68) for those trials where initial focus is classified accurately compared to where it is not classified accurately (0.54). This is unexpected if prosodic initial focus realization has the effect that phrasing cues to constituency are neutralized post-focally. The results suggest, then, that constituency is recoverable from phrasing cues post-focally in Mandarin.

Focus Realization		Constituency Accuracy
A	Accurate	68%
	Inaccurate	54%
B	Accurate	61%
	Inaccurate	63%
C	Accurate	60%
	Inaccurate	56%
W	Accurate	72%
	Inaccurate	65%

**Table 3:** Constituency classification accuracy for each condition of focus realization in the random forest model (Mandarin). Chance level is 50%.

### 3.1.4. Linear mixed-effect model results (Mandarin)

The random forest models help us identify the features that are most informative for the classification of focus and constituency. We used LMMs to test whether the effects of focus and constituency conditions on these features are significant, and in particular whether these two prosodic dimensions interact in their acoustic effects. Results of the model are summarized in **Table 4**, with significant positive and significant negative effects (significant level:  $p < 0.05$ ) marked with + and –, respectively. Non-significant effects are left blank. The complete table of numbers is available in Table S2 in the supplementary materials.

	First syllable (the last name)			Second syllable (the honorific)		
	Max F0	Duration	Intensity	Max F0	Duration	Intensity
(Intercept)	+	+	+	+	+	+
others.vs.noFocus	–	–	–	–	–	–
wide.vs.focus	+	+	–		+	–
preFocus.vs.postFocus	–		–	–		–
noBoundary.vs.bound- aryInternal		+		–	+	–
final.vs.others		+	–		+	+
T4.vs.T1	–	+		+	–	+
position	–		–	–	–	–
preFocus.vs.postFocus: noBoundary.vs.bound- aryInternal					–	
noBoundary.vs.bound- aryInternal: T4.vs.T1						
others.vs.noFocus:T4. vs.T1	+			–		–

**Table 4:** Regression model results for the acoustic cues of both syllables in the Mandarin names

From **Table 4**, non-focused names (for both the last name and the honorific) were reduced compared to wide focus and focused names in all three cues (indicated by the significantly negative effects in the row of *others.vs.noFocus*). To facilitate a better understanding of the model results, we interpret the model estimates within the original feature dimensions below. Max F0 of non-focused last names was estimated to be 40 Hz lower than the wide- or narrow-focused

names, and the reductions for duration and intensity were estimated to be 34 ms and 0.8 dB, respectively. Likewise, for the honorifics, Max F0, duration and intensity were estimated to decrease by 11 Hz, 14 ms and 0.7 dB, respectively.

Focused constituents exhibited higher Max F0 and longer duration, by 17 Hz and 48 ms, respectively, compared to the wide focus case for the last names (effects of *wide.vs.focus*). They were also estimated to be 16 ms longer for the honorific endings. Unexpectedly, the intensity of the name was lower (by 0.5 dB for the last names and by 0.6 dB for the honorifics) in the focus condition than the wide focus condition. Post-focus names (both syllables) had lower F0 (by 63 Hz for the last names and by 47 Hz for the honorifics) and intensity (by 1.8 dB last names and by 2.8 dB for the honorifics) than pre-focus names (effects of *preFocus.vs.postFocus*).

Internal boundaries were expressed mainly by properties of the honorifics (effects of *noBoundary.vs.boundaryInternal*). The honorifics preceding an internal boundary showed longer duration, lower max F0, and lower intensity by 48 ms, 6 Hz, and 0.8 dB, respectively, than those that did not precede any boundary. T1 and T4 names differed significantly in most of the cues, except for intensity of the last names (effects of *T4.vs.T1*), and focus affected T1 and T4 names differently on F0 for both syllables and intensity for the honorifics (effects of *others.vs.noFocus:T4.vs.T1*). However, constituency affected them in the same way (effects of *noBoundary.vs.boundaryInternal:T4.vs.T1*). Even though not of core interest, *position* and *final.vs.others* explained a lot of variation in the models.

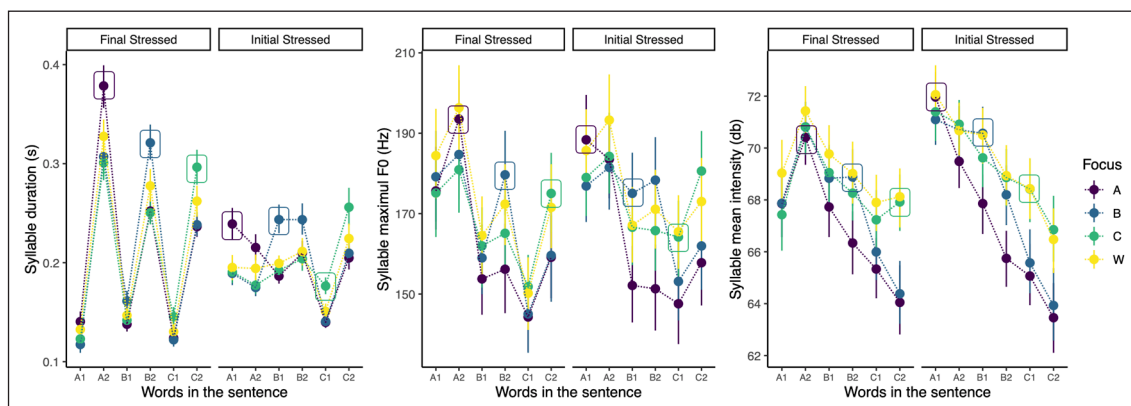
One of our main research questions is the interaction between focus and constituency in their effect on the various acoustic measures (*preFocus.vs.postFocus:noBoundary.vs.boundaryInternal*). These interaction effects were not significant for the acoustic cues other than duration of the honorifics, indicating that the effects of focus and constituency were mostly independent.

## 3.2. Results of English

### 3.2.1. Focus realization (English)

Figure 6 shows the focus effect on the three acoustic cues of the target names, grouped by the names' stress patterns. Like the Mandarin results, both duration and max F0 are informative cues to focus. For the first syllable in the names, duration is, on average, 26 ms longer when on focus compared to the wide focus condition, averaging over constituency and stress pattern. For the second syllable in the names, duration is longer by 35 ms. Max F0 marks the focus mainly through the stressed syllable in the name. Stressed syllables exhibited higher Max F0 by 6 Hz when they are on focus, compared to when they are in the wide focus condition. The post-focal decrease of F0 is observed as well. For example, when A is on focus, the Max F0 of the first syllable of B is 29 Hz lower than that of A, averaging over constituency and stress pattern. The effect of focus on intensity seems to be weaker and is only clear in the C focus condition.

Final-stressed names have larger ranges of variation in duration than initial-stressed names. Specifically, the stress pattern seems to interact with the focus effect. For final-stressed names, when they are focused, longer duration and higher max F0 are observed mainly on the stressed syllable in the names. For initial-stressed names, however, longer duration and higher max F0 are observed on both syllables in the names.

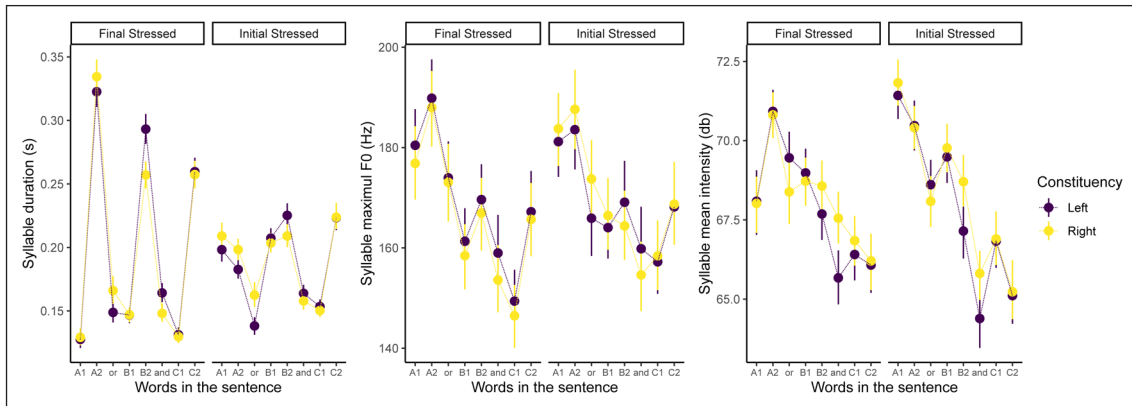


**Figure 6:** Syllable duration, max F0, and intensity of the target names, pooled by focus condition (indicated by colors) and the stress pattern (split by panels) of the English names. Rectangles label the stressed syllable of the focused name for each focus condition. In the legend, A, B, C, and W represent focus on A, B and C and wide focus, respectively. The A1 and A2 on the x-axis indicate the first and second syllables in the first English names, the same for the second (B1, B2) and third (C1, C2) names.

### 3.2.2. Effects of constituency (English)

Figure 7 show the effects of constituency on the three acoustic cues of the coordinate structure, grouped by the stress patterns of the English names. As expected, like the Mandarin results, duration seems to be the most contributive cue to mark the difference in constituency, if it is encoded by phrasing, resulting in final lengthening. Duration of the syllables around the phrasing-induced boundaries were longer than when no boundary was expected. For instance, the duration of B2 is 26 ms longer in the Left branching condition compared to Right branching condition, averaging over focus and stress pattern. Furthermore, intensity was a cue to constituency in English, as well. As expected, intensity was lower at the right edge of constituents, if they precede phrase boundaries. For the Left branching condition, the intensity of B2 is 1.2 dB lower than that in the Right branching condition, averaging over focus and stress pattern.

The max F0 result in English, however, is different from that in Mandarin. For initial-stressed names, syllables around the constituent boundaries have higher max F0, rather than lower, as in Mandarin. As will be discussed later, this could be due to the particular choice of intonation, treating the three consecutive names items on a list, each with a continuation rise. The constituency effect on max F0 is less obvious in the final-stressed names.

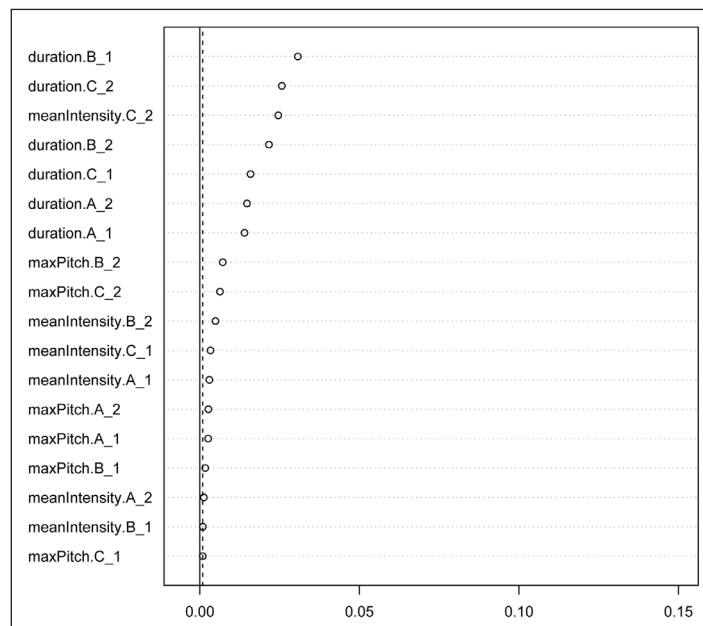


**Figure 7:** Syllable duration, max F0 and intensity for Left and Right branching conditions, grouped by stress pattern of the English names. Panels are the same as in **Figure 6**.

### 3.2.3. Random forest results (English)

#### 3.2.3.1. Random forest model of focus classification (English)

The random forest results of focus classification are shown in **Figure 8**. The most contributive cue to focus seems to be duration (of both syllables in the names), which is different from that in Mandarin, which was max F0. Intensity of the last syllable of name C also contributes substantially. The overall classification accuracy of the random forest model was 51.3% (chance level = 25%). This is lower than the classification in the Mandarin model. The confusion matrix is shown in **Table 5**. As discussed later, the lower accuracy of the classification is most likely because the data quality collected online is poorer, rather than real language differences in prosody.



**Figure 8:** Relative contribution of the features in the random forest model of focus classification (English).

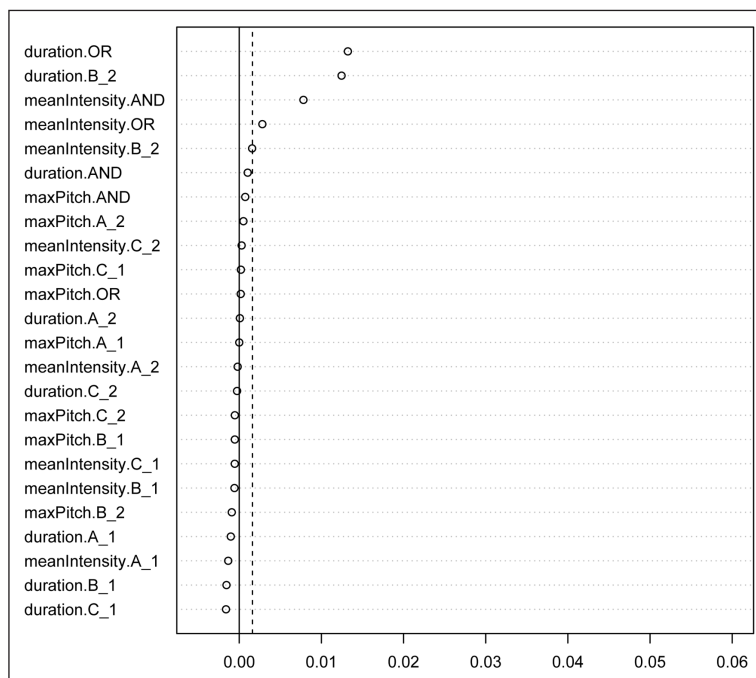


	A	B	C	W	Accuracy
A	160	32	28	47	60%
B	40	159	25	40	60%
C	32	23	156	56	58%
W	53	54	89	71	27%

**Table 5:** Confusion matrix of the random forest model for Focus classification (English). Row labels indicate the intended focus condition. Column labels indicate the classified focus condition. Chance level is 25%.

### 3.2.3.2. Random forest model of constituency classification (English)

The relative ranking of features in the constituency classification of the English data is shown in **Figure 9**. The overall classification accuracy (out-of-bag prediction) was 53%, also lower than that in the Mandarin model, and there are less contributive features than in the Mandarin model. F0 does not contribute to the classification in English, which is different from the Mandarin result. What’s similar to the Mandarin result, though, is that the main cues for constituency are duration and intensity of the pre- and post-boundary syllables as expected, if constituency is encoded by prosodic phrasing.



**Figure 9:** Relative contribution of the features in the random forest models for classifying Constituency (English).

To further analyze the interaction between focus and constituency, the constituency accuracy for each focus condition is shown in **Table 6**. Similar to the Mandarin results, the constituency accuracy of the initial focus condition is not the lowest across all focus type. Rather, it is higher than the B-focus and C-focus conditions. Furthermore, similar to Mandarin, we also observe here that constituency accuracy for the initial focus case is not lower when initial focus is accurately classified than when it is not (56% vs. 53%). This shows that it is not the case that in those productions in which constituency is successfully encoded, participants simply didn't realize initial focus. These results suggest that the post-focus cues to constituency—that is, distinctions in prosodic phrasing—are not neutralized in English, consistent with the finding in Wagner and McAuliffe (2019).

**Table 6** also shows that the constituency classification accuracy of the A-focus condition is quite high. The worst accuracy is observed in the B-focus condition, suggesting that when B is cuing both constituency and focus information, the constituency information is less effectively expressed.

Additionally, we found that the overall constituency classification accuracy in this study is lower than in Wagner and McAuliffe's findings (2019). The overall accuracy is 63% in Mandarin and 53% in English, compared to between 74–79% in Wagner and McAuliffe.

Focus Realization		Constituency Accuracy
A	Accurate	56%
	Inaccurate	53%
B	Accurate	52%
	Inaccurate	45%
C	Accurate	57%
	Inaccurate	43%
W	Accurate	45%
	Inaccurate	62%

**Table 6:** Constituency classification accuracy for each condition of focus realization in the random forest model (English). Chance level is 50%.

### 3.2.4. Linear mixed-effect model results (English)

Linear mixed-effect model results of the acoustic features for both syllables in the English names are shown in **Table 7**. The full table of numbers is in Table S3 in the supplementary materials. As in the Mandarin results, we interpret the model estimates within the original feature dimension to facilitate the understanding of the model results. Compared to wide focus and focused names,

non-focused names (for both syllables in the names) were weaker in all three cues (effects of *others.vs.noFocus*). For the first syllable in the name, Max F0, duration and intensity were estimated to decrease by 9 Hz, 22 ms and 1.5 dB, respectively, when it was not on focus compared to when it is on focus or in the wide focus condition. For the second syllable, the corresponding decreases were estimated to be 14 Hz, 36 ms, and 1.4 dB, respectively. Post-focus names (for both syllables) had lower F0 (by 12 Hz for the first syllable and 14 Hz for the second) than pre-focus names and lower intensity (by 2.3 dB for the first syllable and 4.8 dB for the second) than pre-focus names (effects of *preFocus.vs.postFocus*). The significance of *preFocus.vs.postFocus* was the same as in the Mandarin data, however, as can be found in tables S2 and S3, the amount of the F0 decrease was larger in Mandarin (last name: 68 Hz; honorific: 47 Hz) than in English, suggesting that the post-focus F0 compression is more extreme in Mandarin. Additionally, focused names were longer than wide focus names (by 36 ms for the first syllable and by 52 ms for the second), but had lower mean intensity (by 0.6 dB for the first syllable and by 0.7 dB for the second) than wide focus names (effects of *wide.vs.focus*). This unexpected lower intensity was also observed in Mandarin.

	First syllable			Second syllable		
	Max F0	Duration	Intensity	Max F0	Duration	Intensity
(Intercept)	+	+	+	+	+	+
<i>others.vs.noFocus</i>	-	-	-	-	-	-
<i>wide.vs.focus</i>		+	-		+	-
<i>preFocus.vs.postFocus</i>	-		-	-		-
<i>noBoundary.vs.boundaryInternal</i>					+	-
<i>final.vs.others</i>	-	+	+	-		
<i>FN_S.vs.IN_S</i>	+	+	+		-	-
<i>position</i>	-	+		-	-	-
<i>preFocus.vs.postFocus: noBoundary.vs.boundaryInternal</i>						
<i>noBoundary.vs.boundaryInternal: FN_S.vs.IN_S</i>						
<i>others.vs.noFocus:FN_S.vs.IN_S</i>	-	-	-		+	

**Table 7:** Regression model results for the acoustic cues of both syllables in the English names.

Internal boundaries in English names, as in Mandarin, were expressed mainly by the second syllables of the names (effects of *noBoundary.vs.boundaryInternal*). The second syllables preceding an internal boundary were estimated to be 38 ms longer and 1 dB lower in intensity than those that did not precede any boundary. However, the second syllable showed a significantly higher F0 in Mandarin, whereas such significant F0 effect was not shown in English.

The stress pattern of the English names played a significant role for all the cues of the first syllable and for duration and intensity of the second syllable (effects of *FN\_S.vs.IN\_S*). The stress pattern also modified the effect of focus on (mainly) the first syllable (effects of *others.vs.noFocus:FN\_S.vs.IN\_S*). Furthermore, similar to the Mandarin results, even though not of core interest, *position* and *final.vs.others* explained considerable variation in the model.

Importantly, **Table 7** shows that the interaction between focus and constituency (*preFocus.vs.postFocus:noBoundary.vs.boundaryInternal*) was not significant in any investigated features, indicating that the effects of focus and constituency were realized independently in English, consistent with the results in Wagner and McAuliffe (2019).

### 3.3. Results of LMM models for both languages combined

In sections 3.1 and 3.2, we reported LMM results of each language separately, and the result showed that in both languages the focus and constituency did not interact with each other in their acoustic realizations, other than in the duration of the honorific suffix in Mandarin. From these separate models, we cannot draw any conclusions whether the two languages differ (Nieuwenhuis et al., 2011). Using data from English and Mandarin, we fit an LMM with a three-way interaction term among language, focus, and constituency to see how similar or different the two languages are.

Since the lexical patterns of English and Mandarin are different, we used the subset of initial stressed names in English and the subset of T1 names in Mandarin. All the Mandarin names in the experiment were initial-stressed, making initial-stressed English names a better match. Moreover, there exists a downstep effect in consecutive T4s in an utterance, so avoiding the Mandarin T4 names makes the mixed data more consistent in terms of the tonal effect. This also eliminated the terms *T4.vs.T1* and *FN\_S.vs.IN\_S* in the models. An additional predictor for language was coded as below:

- EN.vs.MD: The value is 1 if the syllable is in a Mandarin name. The value is -1 if it is in an English name.

We fit an LMM, which included a three-way interaction term *noBoundary.vs.boundaryInternal\*preFocus.vs.postFocus\*EN.vs.MD*, and two two-way interaction terms, *noBoundary.vs.boundaryInternal\*EN.vs.MD* and *preFocus.vs.postFocus\*EN.vs.MD*, for each of the

acoustic cues and for both syllables on the mixed data (initial stressed English names and T1 Mandarin names). As we hypothesized that Mandarin and English show differences in using the acoustic cues such as duration and F0 in marking focus and constituency, we would expect the above two two-way effects to be significant for some of the three cues.

The model results are summarized in Table S4 in the supplementary materials. The three-way interaction effect is non-significant in all six models, suggesting that the two languages do not significantly differ with respect to the focus-constituency interaction. Given that none of the models showed significant *preFocus.vs.postFocus:noBoundary.vs.boundaryInternal*, we confirmed that in both Mandarin and English, the post-focal cue to constituency is not neutralized, suggesting that phrasing distinctions in the post-focal domain are intact. Furthermore, we found that English and Mandarin differ significantly in the effects of focus (*preFocus.vs.postFocus:EN.vs.MD*) on Max F0. In the post-focal domain F0 is lower, but in Mandarin it is lowered more than in English, and this is shown on both syllables (effects of *preFocus.vs.postFocus:EN.vs.MD*). For the first syllable, the post-focal max F0 lowering in Mandarin is 31 Hz greater than in English, and for the second syllable, it is 44 Hz greater than in English.

In terms of the effect of constituency, the combined model showed that the two languages differ significantly only in the max F0 of the second syllable (effects of *noBoundary.vs.boundaryInternal:EN.vs.MD*). This latter result is consistent with the seemingly divergent role of max F0 on the accented syllable in the separate regression models in the two languages: The max F0 of the name's second syllable in the internal boundary condition is significantly higher in English but significantly lower in Mandarin than in the no boundary condition.

## 4. Discussion

The current study investigated whether focus and constituency interact with each other in Mandarin and in English, through a production experiment using coordinated names. Specifically, we focused on the question of whether post-focal cues to constituency will be neutralized or not. Results revealed that they are not, suggesting that prosodic phrasing distinctions remain intact post-focally both in Mandarin and English, and not neutralized. This finding is consistent with some previous results in English and German. Since Mandarin and English have different lexical prosodic patterns (lexical tone vs. lexical stress), we also investigated to what extent the two languages differ in their prosodic correlates of focus and constituency. These similarities and differences are discussed in the next sections.

### 4.1. Focus realization in Mandarin and in English

In both English and Mandarin, when a bi-syllabic name is focused, the focus information is mainly expressed through strengthening the most prominent syllable in the name (the stressed

syllable in English and the first syllable in Mandarin), although the less prominent syllable in the name gets strengthened, as well. These results are consistent with the general conclusion that F0, duration, and intensity all contribute to focus marking (Cooper et al., 1985; as reviewed in Cole, 2015) for both English and Mandarin.

What differs in the two languages, however, is the relative importance of the three cues in marking focus, as shown by the random forests: F0 makes the most contribution for Mandarin, and duration makes the most contribution for English. This result is compatible with the idea that Mandarin is a language that marks focus predominantly by modulating pitch range (e.g., Kügler & Calhoun, 2020), but given the relevance of all cues in both languages, this might not be a crisp typological difference but rather a gradient difference in the weight of particular cues.

This result for English in our study differs from Wagner and McAuliffe (2019), where F0 was the most important cue to focus. The discrepancy could be because of the different paradigms used in these two experiments. In the present study, the target sentence was elicited in dialogues with prerecorded contexts, whereas in Wagner and McAuliffe there was no prerecorded interlocutor. Furthermore, participants had to formulate the sentence themselves, and images and distances were used to elicit particular constituent structures rather than commas. Although participants might not behave as spontaneously as they do in a daily conversation, as the texts of the materials were inevitably presented to them on the screen, our task was more similar to a conversation than Wagner and McAuliffe's task, which was overall more like a reading task. Read speech is quite different from spontaneous speech, including the realization of focus (e.g., Koopmans-Van Beinum, 1992; Laan, 1997; see Wagner et al., 2015, for a review), which may partly account for different roles that F0 played in marking focus in English in the two studies.

The random forest ranks the contribution of local cues which are from the focused names, but focus is also cued by features outside of the focal domain. F0 can also mark focus through post-focal compression (Xu, 1999), and our results are compatible with this interpretation. Compared to pre-focus names, post-focus names had lower F0 in both Mandarin and English. However, the amount of post-focal F0 decrease was greater in Mandarin than in English. This may be due to Mandarin being considered a pitch-range-based focus marking system (Kügler & Calhoun, 2020), which has shown both compressed and lowered pitch range in the post-focal domain (Xu, 1999), whereas English only shows the general lowering of F0 post-focally. It should be mentioned that the pitch lowering, or the 'post-focal compression' observed in this study, is determined based on the parameter of maximum F0. It's noteworthy that mean F0 or F0 range, which are parameters offering a more nuanced description of F0 change or F0 range compression, were not used. This methodological choice may introduce bias in the results, potentially towards a smaller observed effect. Other than F0, the mean intensity in the post-focal domain was also smaller than the pre focal domain in both languages. From these results, post-focal weakening happens in both F0

and intensity dimensions but not duration for both languages, and a larger degree of post-focal compression was observed in Mandarin.

Additionally, focus encoding seems to be more consistent in Mandarin than in English. The random forest results found that the focus classification accuracy in Mandarin was higher than in English (both the English results in this study and in Wagner and McAuliffe, 2019). In both languages, there was more confusion between wide focus and final focus in the random forest models, confirming the previous finding that, in principle, the nuclear prominence is located by default on the rightmost word in the prosodic phrase and that wide focus is realized phonetically in a way similar to final focus, cross-linguistically (Chen et al., 2016; Ladd, 2008). Tables 2 and 5 show that the confusion between wide focus and final focus was larger in English than in Mandarin, however.

## 4.2. Cues to constituency in Mandarin and English

The results show that in both Mandarin and English, constituency is mostly marked by the syllable before and after the constituent boundary, i.e., the second syllable in the names and the coordinators as expected, if the phonological means to encode constituency is prosodic phrasing. The random forest models found that the phonetic realization of coordinators make a larger contribution to encoding constituency in Mandarin than in English. In both languages, the pre-boundary syllable is longer and has lower intensity. These patterns are consistent with what Wagner and McAuliffe (2019) found for English, and Poschmann and Wagner (2016) for German.

We also found that in both languages the penultimate syllable before the boundary, i.e., for both internal and final boundary in Mandarin and for final boundary in English, the first syllable in the pre-boundary name is lengthened. Studies have found that the pre-boundary lengthening is not restricted to the last syllable before the boundary, and the range of syllables lengthened can be language dependent (Cho, 2015; Cho & McQueen, 2005). Results in the present study showed that in both languages the range of pre-boundary lengthening is no less than two syllables, and the closer it is to the phrasing, the more it gets lengthened. This is compatible with the  $\pi$ -gesture model that boundary effect is strongest at the boundary and will decrease with distance from it (Byrd & Saltzman, 2003), and the replications of this prediction in the articulatory phonology literature (Krivokapić, 2020). Additionally, in contrast to duration, for both languages the intensity lowering is limited to the last syllable before the boundary and does not persist to the penultimate syllable.

So far, we have discussed many similarities between English and Mandarin in prosodically marking constituency. One difference we observed in this study is the role of F0. In the separate models, we observed that the syllable before the internal boundary had lower max F0 in Mandarin but did not show significant max F0 difference in English, which is also confirmed by the significant interaction effect in the joint model. These results seem to show that the role of max F0 in boundary marking differs in Mandarin and in English.



This could be due to differences in the preferred intonation between the two languages. Through listening to productions from each participant, we noticed that English participants, but not Mandarin participants, sometimes used a listing intonation, i.e., F0 rises across the first two names to indicate continuation and falls on the final name to signal the end of the list. Three of the participants used such intonation frequently and others used it less frequently. This type of list intonation was also found in the lab speech of native Spanish speakers in MacLeod and Di Lonardo Burr (2022). Hence, the rising F0 in English speakers' productions, occurring on the first and second names in the phrase, may align well with the internal boundary. Although in a subset of the data participants may use a falling F0 to mark an internal boundary; after mixing with these list intonations, we may not be able to observe any overall effect on F0. It's also possible that in another subset of the data, participants phonologically chose to use a rising F0 to mark the internal boundary. However, further investigations are needed to distinguish this from the listing intonation, and future studies can explore the factors influencing the choice between rising or falling F0 to mark an internal boundary. Such intonational variation may also explain why the contribution of F0 to marking prosodic boundaries is often found to be limited (Seoul Korean: Jeon & Nolan, 2013; English: Lehiste, 1973; Wagner & McAuliffe, 2019). One reliable effect of F0 we found was that it was lower in final position as expected, if there is a final lowering effect (Lieberman & Pierrehumber 1984).

Compared to Wagner and McAuliffe (2019), the random forest models in this study showed lower accuracies of constituency classification. This could be due to the different ways of manipulating constituency in the particular tasks used. Wagner and McAuliffe showed participants the target sentence with commas indicating the phrasing condition, whereas in this study participants relied on the context including the configuration of the pictures. Commas may be taken by listeners as encoding the presence of a prosodic juncture, in addition to a syntactic one, potentially resulting in more consistent prosodic encoding. One explanation is that syntactic constituent structure is only optionally conveyed, and commas increase the likelihood that it is. However, this does not explain the difference in constituency accuracy between Mandarin and English observed here, with higher accuracy in the random forest for constituency in Mandarin.

### **4.3. The interaction of focus and constituency in Mandarin and English**

The results in this study suggested that in both English and Mandarin, post-focal cues to constituency are not always neutralized. In both languages, we found the classification accuracy for constituency in the initial focus condition was high. In fact, the constituency classification accuracy was even higher when the focus classification for initial focus was accurate, confirming that it is not the case that constituency was only conveyed when focus was not realized. This is compatible with the idea that the prosodic correlates of focus and phrasing should be represented separately, perhaps via a metrical representation that orthogonally encodes both. An orthogonal representation of prominence and phrasing is also consistent with the non-interaction between

focus and constituency in the LMMs for English. This independence between the effects of focus and constituency is consistent with previous evidence in English and German (Kügler & Féry, 2017; Norcliffe & Jaeger, 2005; Wagner & McAuliffe, 2019).

In Mandarin, the focus-syntax interaction effect in the duration LMM of the Mandarin honorific syllables was significant. This suggests that in Mandarin, in contrast to English, focus and constituency interact in their effect on duration. However, this difference between the two languages could not be confirmed in models based on the combined data. None of the three-way interaction effects in the LMMs on the combined data was significant, indicating no differences in how English and Mandarin behave in terms of the focus-phrasing interaction.

These results are compatible with additive models of sentence prosody, which assume that different communicative functions can be encoded in an additive way, but it is also compatible with metrical models that cleanly distinguish the representation of prominence and phrasing (see discussion in Wagner and McAuliffe, 2019).

#### **4.4. The role of intensity in focus and phrasing realizations**

Compared to F0 and duration, the role of intensity in encoding prosodic phrasing has received less attention. Wagner and McAuliffe (2019) found that intensity was lower in pre-boundary positions in English. In Mandarin it was hypothesized in Wang, Xu et al. (2018) that intensity might cue phrasing in Mandarin. Results in this study confirmed this hypothesis in Mandarin and replicated the finding in Wagner and McAuliffe for English—in both languages intensity was lower for the pre-boundary syllable (but not for the penultimate syllable).

For focus marking, this study compared the difference of both syllables receiving focus (either alone or part of wide focus) i.e., other vs. no-focus and focus vs. wide focus. For other vs. no-focus, consistent with the general finding of focus marking (Fletcher, 2010; Ladd, 2008), intensity was lower for the syllables without focus than those receiving wide or narrow focus. For focus vs. wide focus, however, contrary to the general finding, we found intensity was lower for focused syllables than syllables in the wide focus condition. The varied roles of intensity shown in the two comparisons also suggested the internal difference between the other vs. no-focus condition, and the focus vs. wide focus condition.

In addition, as mentioned in Section 4.1, similar to the post-focus compression of F0 (Xu, 1999), we found post-focus weakening of intensity in both languages and for both syllables in the name.

#### **4.5. Focus-induced prominence vs. phrasing-induced prominence**

When analyzing whether initial focus would inhibit the realization of cues to constituent structure, we examined the constituency accuracies for each focus condition (Tables 3 and 6) and found that the lowest constituency accuracy was for B-focus conditions rather than initial

focus conditions. This poorer constituency accuracy of the B-focus condition could be due to the acoustic similarity between phrase-induced prominence and focus-induced prominence.

Previous studies found that words or syllables in phrase final positions are perceived as perceptually more prominent than in phrase internal positions (Bishop et al., 2020; Cole et al., 2010; Jagdfeld & Baumann, 2011). Similarly, production results in this study also showed both the focused names and the pre-boundary names have longer duration. Lengthening is thus an overlapped cue for both focus and constituency. B in the coordinated structure is in such a special position that both its focus and phrasal boundary were realized utterance-internally. This restriction would further increase the similarities between the focus-induced prominence and prominence to syntactically motivated phrasing on B (which occurs phrase-finally), versus A and C.

To seek more evidence, we then analyzed the breakdown of predictions from the random forest model of focus. We found that in English, of all the samples that were predicted as B focus, 67% of them were predicted as Left-branching in the random forest model of constituency classification. The percentage in the equivalent Mandarin data was 59%, lower than in English but also above the default value 50%. The result—that most of the predicted B-focus productions were predicted as Left-branching—suggests that the focused name B was acoustically similar to the phrase final names, confirming the acoustic similarities between focus-induced prominence and phrasing-related prominence.

#### **4.6. Effects of Mandarin tone on prosodic cues to focus and constituency**

We included two tones (T1 and T4) in the Mandarin materials and two stress patterns (initial-stressed and final-stressed) in the English materials. Since focus is mainly expressed through the stressed syllable as discussed earlier, it is more intuitive that the stress pattern of the English names affects the focus marking. In the Mandarin materials, however, the prominence pattern was identical and the only difference was the tone type. Therefore, in this section we focus on the role of tone and its interaction with focus marking in Mandarin.

Results of this study showed that the overall duration of last names with T1 were longer than last names with T4, and the max F0 of last names of T4 was higher than last names with T1. These results reflect the intrinsic F0 and duration values of Mandarin T1 and T4 (Xu, 2005). In addition, we also observed that honorifics in the T4 names had significantly lower F0 than honorifics in the T1 names, reflecting the downstep effect triggered by the low targets in the T4 syllables (Shih, 1988).

Furthermore, names with T1 and T4 showed different patterns of focus marking (from term *others.vs.noFocus:T4.vs.T1*), which was also found in some previous studies (Wang et al., 2020; Zhang et al., 2021). Last names with T4 had larger F0 difference between focused (including wide focus) and non-focused conditions than last names with T1. This suggests that T4 relies more on

the F0 dimension in marking focus than T1. Similar findings were shown in Zhang et al. (2021), where mean F0 exhibited a larger effect size in the prominence perception of T4 compared to T1. One possible reason for the greater involvement of F0 in T4 is that T4 is intrinsically shorter than T1 (Xu, 2005). The temporal constraints may impose stronger articulatory demands on T4 to achieve the F0 targets correctly and quickly. In contrast, T1, having weaker temporal constraints and lacking other level tone competitors in Mandarin, allows more flexible F0 realizations. This discrepancy can make F0 a more usable cue for T4 than T1 in marking focus. In contrast to the last names, the honorific in the T4 names have smaller F0 difference between focused and non-focused conditions than the honorifics in T1 names, indicating that the focused last names with T4 suppressed the following syllable in terms of the F0 range. Altogether, these results confirmed that focus marking is also modified by syllable-level linguistic properties.

## 5. Conclusion

In this study we investigated the effect of focus and constituent structure on sentence prosody in Mandarin and in English through a production study in a dialogue task with a prerecorded interlocutor. Contrary to some accounts of sentence prosody, syntactically-motivated phrasing distinctions in the post-focal domain were found to persist rather than be neutralized in both languages.

Regarding focus, we observed that in both Mandarin and English, focus was expressed mainly through cues on the most prominent syllable in a bi-syllabic word (stressed syllable for English, first syllable for Mandarin), although the unstressed syllable exhibited some degree of focus-related effects. Both languages showed evidence of post-focal compression in F0. Interestingly, F0 contributed more to classification of focus in Mandarin than in English, compatible with the idea that Mandarin, but not English, marks focus via pitch range modulation (Kügler & Calhoun, 2020). However, the relevance of other cues such as intensity and duration in both languages suggests that there is no crisp difference between these languages, but rather a difference in degree. Furthermore, we found that the acoustic marking of focus interacted with lexical tone type in Mandarin and stress pattern in English. In general, focus realization was very similar for the two languages, despite the fact that English and Mandarin are sometimes classified as having different focus marking systems (e.g., Kügler and Calhoun 2020).

We found that, as expected, constituency was marked in both languages by the syllables before and following a constituency edge, if constituency is encoded via prosodic phrasing. Pre-boundary syllables had longer duration and lower intensity. The F0-correlate of constituency was less clear, and future studies will be needed to investigate this further. The effect of constituency was shown to be roughly similar in Mandarin and English. For example, lexical tone choice in Mandarin did not interact with phrasing marking, nor did stress pattern in English.

We should also note that there are number of limitations in the current study. We cannot rule out the possibility that some of the results were partially due to the fact that the English data was collected online whereas the Mandarin data was collected in person. Although the use of an automatic forced aligner and our acoustic parameter extraction method has also been used with success in earlier studies such as van der Klok (2018) and Wagner and McAuliffe (2019), it is possible that this purely quantitative approach misses out on important phonological distinctions. Furthermore, in the Mandarin names the second syllable ('ge') was controlled, whereas in the English names the second syllable varied. This discrepancy may have influenced the comparison between the two languages.

This study also informs our understanding of the typology of focus realization. On the one hand, the greater effect of post-focal compression of F0 in Mandarin than in English is compatible with the claim in Kügler and Calhoun (2020) that English is a stress-based-focus-marking language and Mandarin is a pitch-range-based-focus-marking language. On the other hand, the languages are otherwise very similar in how they mark focus, suggesting that there may not be a clean typological distinction here. The high degree of similarity in both focus and constituency realization suggests that the two languages are relatively similar with respect to how their sentence prosody encodes these two dimensions. The results underline the importance of directly comparing languages with similar methodologies, rather than inferring typological differences from separate studies that use very different materials and methodologies, as is often done.

---

## Additional file

The additional file for this article can be found as follows:

- **Supplementary Materials.** Tables S1 to Table S4. DOI: <https://doi.org/10.16995/labphon.9704.s1>

## Acknowledgements

We thank the audience at the 1st conference of TAI (Tone and Intonation) in Dec 2021. This work was funded by SSHRC grant #435-2020-1140 to Meghan Clayards, NSERC grant #RGPIN-2018-06153 to Michael Wagner, and fellowships from CSC and FRQSC to Wei Zhang.

## Competing interests

The authors have no competing interests to declare.

---

## References

- Bailly, G., & Holm, B. (2005). SFC: A trainable prosodic model. *Speech Communication*, 46(3–4), 348–364. <https://doi.org/10.1016/j.specom.2005.04.008>
- Baltazani, M., & Jun, S.-A. (1999). Focus and topic intonation in Greek. *Focus*, 800(900), 1000.
- Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. (R package, Version 1.1-7). [Computer software]. <https://doi:10.18637/jss.v067.i01>
- Beckman, M. E., & Ayers, G. (1997). Guidelines for ToBI labelling. *The OSU Research Foundation*, 3, 30.
- Bishop, J., Kuo, G., & Kim, B. (2020). Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from Rapid Prosody Transcription. *Journal of Phonetics*, 82, 100977. <https://doi.org/10.1016/j.wocn.2020.100977>
- Büiring, D. (2009). Towards a typology of focus realization. In M. Zimmermann & C. Féry (Eds.), *Information Structure: Theoretical, Typological, and Experimental Perspectives*, 177–205. Oxford Academic. <https://doi.org/10.1093/acprof:oso/9780199570959.003.0008>
- Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, 31(2), 149–180. [https://doi.org/10.1016/S0095-4470\(02\)00085-2](https://doi.org/10.1016/S0095-4470(02)00085-2)
- Calhoun, S. (2006). *Information structure and the prosodic structure of English: A probabilistic relationship*. [Doctoral dissertation, University of Edinburgh]. <http://hdl.handle.net/1842/8120>
- Calhoun, S., Wollum, E., & Kruse Va'ai, E. (2021). Prosodic prominence and focus: Expectation affects interpretation in Samoan and English. *Language and Speech*, 64(2), 346–380. <https://doi.org/10.1177/0023830919890362>

- Cambier-Langeveld, T. (1997). The domain of final lengthening in the production of Dutch. *Linguistics in the Netherlands*, 14(1), 13–24. <https://doi.org/10.1075/avt.14.04cam>
- Cambier-Langeveld, T. (1999). The Interaction between final lengthening and accentual Lengthening: Dutch versus English. *Linguistics in the Netherlands*, 16(1), 13–15. <https://doi.org/10.1075/avt.16.04cam>
- Chao, Y. R. (1965). *A grammar of spoken Chinese*. University of California Press.
- Chen, Y. (2010). Post-focus F0 compression—Now you see it, now you don't. *Journal of Phonetics*, 38(4), 517–525. <https://doi.org/10.1016/j.wocn.2010.06.004>
- Chen, Y., & Gussenhoven, C. (2008). Emphasis and tonal implementation in Standard Chinese. *Journal of Phonetics*, 36(4), 724–746. <https://doi.org/10.1016/j.wocn.2008.06.003>
- Chen, Y., Lee, P. P.-L., & Pan, H. (2016). Topic and focus marking in Chinese. In C. Féry & S. Ishihara (Eds.), *The Oxford Handbook of Information Structure*. Oxford Academic. <https://doi.org/10.1093/oxfordhb/9780199642670.013.34>
- Cho, T. (2015). Language Effects on Timing at the Segmental and Suprasegmental Levels. In M. A. Redford (Ed.), *The Handbook of Speech Production* (pp. 505–529). John Wiley & Sons. <https://doi.org/10.1002/9781118584156.ch22>
- Cho, T., & Keating, P. A. (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics*, 29(2), 155–190. <https://doi.org/10.1006/jpho.2001.0131>
- Cho, T., & McQueen, J. M. (2005). Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics*, 33(2), 121–157. <https://doi.org/10.1016/j.wocn.2005.01.001>
- Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience*, 30(1–2), 1–31. <https://doi.org/10.1080/23273798.2014.963130>
- Cole, J., Mo, Y., & Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, 1(2). <https://doi.org/10.1515/labphon.2010.022>
- Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question–answer contexts. *The Journal of the Acoustical Society of America*, 77(6), 2142–2156. <https://doi.org/10.1121/1.392372>
- Cruttenden, A. (1994). Phonetic and prosodic aspects of baby talk. In C. Gallaway & B. J. Richards (Eds.), *Input and Interaction in Language Acquisition* (pp. 35–152). Cambridge University Press. <https://doi.org/10.1017/CBO9780511620690.008>
- De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1), 1–12.
- De Pijper, J. R., & Sanderman, A. A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *The Journal of the Acoustical Society of America*, 96(4), 2037–2047. <https://doi.org/10.1121/1.410145>
- DiCanio, C., Benn, J., & Castillo García, R. (2018). The phonetics of information structure in Yoloxóchitl Mixtec. *Journal of Phonetics*, 68, 50–68. <https://doi.org/10.1016/j.wocn.2018.03.001>



- Féry, C. (2014). Final compression in French as a phrasal phenomenon. In S. Katz Bourns & L. L. Myers (Eds.), *Perspectives on Linguistic Structure and Context: Studies in Honour of Knud Lambrecht* (pp. 133–156). John Benjamins. <https://doi.org/10.1075/pbns.244.07fer>
- Féry, C., & Kügler, F. (2008). Pitch accent scaling on given, new and focused constituents in German. *Journal of Phonetics*, 36(4), 680–703. <https://doi.org/10.1016/j.wocn.2008.05.001>
- Fletcher, J. (2010). The prosody of speech: Timing and rhythm. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences*, (pp. 521–602). Wiley. <https://doi.org/10.1002/9781444317251.ch15>
- Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, 101(6), 3728–3740. <https://doi.org/10.1121/1.418332>
- Fujisaki, H. (1983). Dynamic characteristics of voice fundamental frequency in speech and singing. In P. F. MacNeilage (Ed.), *The Production of Speech* (pp. 39–55). Springer. [https://doi.org/10.1007/978-1-4613-8202-7\\_3](https://doi.org/10.1007/978-1-4613-8202-7_3)
- Gollrad, A. (2013). Prosodic cue weighting in sentence comprehension: Processing German case ambiguous structures. [Doctoral dissertation, University of Potsdam].
- Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge University Press. <https://doi-org/10.1017/CBO9780511616983>
- Hamlaoui, F., Žygis, M., Engelmann, J., & Wagner, M. (2019). Acoustic correlates of focus marking in Czech and Polish. *Language and Speech*, 62(2), 358–377. <https://doi.org/10.1177/0023830918773536>
- Hirst, D., & Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix*, 15, 75–85
- Hothorn, T., Hornik, K., Strobl, C., & Zeileis, A. (2010). *Party: A laboratory for recursive partytioning*. The Comprehensive R Archive Network. <https://doi.org/10.32614/CRAN.package.party>
- Jagdfeld, N., & Baumann, S. (2011). Order Effects on the Perception of Relative Prominence. *Proceedings 17th International Congress of Phonetic Sciences*. Hong Kong, China (pp. 958–961). ICPhS.
- Jeon, H.-S., & Nolan, F. (2013). The role of pitch and timing cues in the perception of phrasal grouping in Seoul Korean. *The Journal of the Acoustical Society of America*, 133(5), 3039–3049. <https://doi.org/10.1121/1.4798663>
- Jeon, H.-S., & Nolan, F. (2017). Prosodic marking of narrow focus in Seoul Korean. *Laboratory Phonology*, 8(1), 1–30. <https://doi.org/10.5334/labphon.48>
- Jun, S.-A., & Lee, H.-J. (1998). Phonetic and phonological markers of contrastive focus in Korean. *Fifth International Conference on Spoken Language Processing*. Sydney, Australia. <https://doi.org/10.21437/ICSLP.1998-151>
- Kalinowski, C. (2015). A typology of morphosyntactic encoding of focus in African languages. (Publication No. 3725934). [Doctoral dissertation, State University of New York, Buffalo]. ProQuest Dissertations & Theses.

- Kanerva, J. M. (1991). Focus and phrasing in Chichewa phonology. [Doctoral dissertation, Stanford University].
- Kim, H., Yoon, T., Cole, J., & Hasegawa-Johnson, M. (2006). Acoustic differentiation of L-and LL% in switchboard and radio news speech. In R. Hoffmann, H. Mixdorff (Eds.), *3rd International Conference on Speech Prosody (Proceedings of the International Conference on Speech Prosody)* (pp. 214–217). International Speech Communication Association.
- Kim, S., Kim, J., & Cho, T. (2018). Prosodic-structural modulation of stop voicing contrast along the VOT continuum in trochaic and iambic words in American English. *Journal of Phonetics*, *71*, 65–80. <https://doi.org/10.1016/j.wocn.2018.07.004>
- Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, *3*(3), 129–140. [https://doi.org/10.1016/S0095-4470\(19\)31360-9](https://doi.org/10.1016/S0095-4470(19)31360-9)
- Kochanski, G., & Shih, C. (2003). Prosody modeling with soft templates. *Speech Communication*, *39*(3–4), 311–352. [https://doi.org/10.1016/S0167-6393\(02\)00047-X](https://doi.org/10.1016/S0167-6393(02)00047-X)
- Koopmans-Van Beinum, F. J. (1992). The role of focus words in natural and in synthetic continuous speech: Acoustic aspects. *Speech Communication*, *11*(4–5), 439–452. [https://doi.org/10.1016/0167-6393\(92\)90049-D](https://doi.org/10.1016/0167-6393(92)90049-D)
- Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, *55*(3–4), 243–276.
- Krivokapić, J. (2020). Prosody in articulatory phonology. In J. Barnes & S. Shattuck-Hufnagel (Eds.), *Prosodic Theory and Practice* (pp. 213–236). MIT Press. <https://doi.org/10.7551/mitpress/10413.001.0001>
- Kügler, F. (2020). Post-focal compression as a prosodic cue for focus perception in Hindi. *Journal of South Asian Linguistics*, *10*, 38–59.
- Kügler, F., & Calhoun, S. (2020). Prosodic Encoding of Information Structure. In C. Gussenhoven & A. Chen (Eds.), *The Oxford Handbook of Language Prosody*. Oxford Academic. <https://doi.org/10.1093/oxfordhb/9780198832232.013.30>
- Kügler, F., & Féry, C. (2017). Postfocal Downstep in German. *Language and Speech*, *60*(2), 260–288. <https://doi.org/10.1177/0023830916647204>
- Kügler, F., & Genzel, S. (2012). On the prosodic expression of pragmatic prominence: The case of pitch register lowering in Akan. *Language and Speech*, *55*(3), 331–359. <https://doi.org/10.1177/0023830911422182>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Laan, G. P. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication*, *22*(1), 43–65. [https://doi.org/10.1016/S0167-6393\(97\)00012-5](https://doi.org/10.1016/S0167-6393(97)00012-5)
- Ladd, D. R. (1988). Declination “reset” and the hierarchical organization of utterances. *The Journal of the Acoustical Society of America*, *84*(2), 530–544. <https://doi.org/10.1121/1.396830>

- Ladd, D. R. (1990). Intonation: Emotion vs. grammar. *Language*, 66(4), 806–816. <https://doi.org/10.2307/414730>
- Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511808814>
- Lehiste, I. (1973). Rhythmic units and syntactic units in production and perception. *The Journal of the Acoustical Society of America*, 54(5), 1228–1234. <https://doi.org/10.1121/1.1914379>
- Li, A. (2002). Chinese prosody and prosodic labeling of spontaneous speech. *Proceedings of Speech Prosody 2002*. Aix-en-Provence, France. <https://doi.org/10.21437/SpeechProsody.2002-6>
- Lin, M. (1999). Breaks and prosodic phrases in the utterances of Standard Chinese. *14th International Congress of Phonetic Sciences* (pp. 2109–2112). ICPhS Archive. [https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14\\_2109.pdf](https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_2109.pdf)
- Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, 35(11), 1773–1781. <https://doi.org/10.1121/1.1918816>
- MacLeod, B., & Di Lonardo Burr, S. M. (2022). Phonetic imitation of the acoustic realization of stress in Spanish: Production and perception. *Journal of Phonetics*, 92, 101139. <https://doi.org/10.1016/j.wocn.2022.101139>
- McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using Kaldi. *Interspeech 2017* (pp. 498–502). ICSA Archive. <https://doi.org/10.21437/Interspeech>.
- Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, 14(9), 1105–1107. <https://doi.org/10.1038/nn.2886>
- Norcliffe, E., & Jaeger, T. F. (2005). Accent-free prosodic phrases? Accents and phrasing in the post-nuclear domain. *Proceedings of Interspeech 2005*.
- Paschen, L., Fuchs, S., & Seifart, F. (2022). Final lengthening and vowel length in 25 languages. *Journal of Phonetics*, 94, 101179. <https://doi.org/10.1016/j.wocn.2022.101179>
- Pierrehumbert, J. (1979). The perception of fundamental frequency declination. *The Journal of the Acoustical Society of America*, 66(2), 363–369. <https://doi.org/10.1121/1.383670>
- Pierrehumbert, J., & Talkin, D. (1992). Lenition of/h/and glottal stop. In G. J. Docherty & D. R. Ladd (Eds.), *Papers in laboratory Phonology II: Gesture, segment, prosody* (pp. 90–117). Cambridge University Press.
- Poschmann, C., & Wagner, M. (2016). Relative clause extraposition and prosody in German. *Natural Language & Linguistic Theory*, 34(3), 1021–1066.
- R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Royer, J. (2022). Prosody as syntactic evidence. *Natural Language & Linguistic Theory*, 40(1), 239–284.
- Selkirk, E. O. (1986). *Phonology and syntax: The relationship between sound and structure*. MIT Press.

- Seo, J., Kim, S., Kubozono, H., & Cho, T. (2019). Preboundary lengthening in Japanese: To what extent do lexical pitch accent and moraic structure matter? *The Journal of the Acoustical Society of America*, 146(3), 1817–1823. <https://doi.org/10.1121/1.5122191>
- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. *Proceedings of the Second International Conference on Spoken Language Processing*, (pp. 867–870). ISCA Archive. <https://doi.org/10.21437/ICSLP.1992-260>
- Steedman, M. (1991). Structure and intonation. *Language*, 67(2), 260–296. <https://doi.org/10.2307/415107>
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14(4), 323. <https://doi.org/10.1037/a0016973>
- Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *The Journal of the Acoustical Society of America*, 101(1), 514–521. <https://doi.org/10.1121/1.418114>
- Taylor, P. (2000). Analysis and synthesis of intonation using the tilt model. *The Journal of the Acoustical Society of America*, 107(3), 1697–1714. <https://doi.org/10.1121/1.428453>
- Truckenbrodt, H. (2002). Upstep and embedded register levels. *Phonology*, 19(1), 77–120. <https://doi.org/10.1017/S095267570200427X>
- Tsai, K., & Katsika, A. (2020). Pitch accent and phrase boundaries: Kinematic evidence from Japanese. *Proceedings of the 10th International Conference on Speech Prosody 2020*. ISCA Archive. <https://doi.org/10.21437/SpeechProsody.2020-44>
- Tseng, C., Su, Z., & Lee, L. (2009). Mandarin spontaneous narrative planning-prosodic evidence from National Taiwan University lecture corpus. *Proceedings of Interspeech 2009*. ISCA Archive. <https://doi.org/10.21437/Interspeech.2009-745>
- Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, 35(4), 445–472. <https://doi.org/10.1016/j.wocn.2006.12.001>
- Van Santen, J. P. H., & Möbius, B. (2000). A Quantitative Model of Fo Generation and Alignment. In A. Botinis (Ed.), *Intonation: Analysis, Modelling and Technology* (pp. 269–288). Springer. [https://doi.org/10.1007/978-94-011-4317-2\\_12](https://doi.org/10.1007/978-94-011-4317-2_12)
- Vander Klok, J. M., Goad, H., & Wagner, M. (2018). Prosodic focus in English vs. French: A scope account. *Glossa: A Journal of General Linguistics*, 3(1). <https://doi.org/10.5334/gjgl.172>
- Wagner, M. (2005). Prosody and recursion. [Doctoral dissertation, Massachusetts Institute of Technology].
- Wagner, M. (2021). Prosody lab experimenter. [Computer software]. Prosodylab. <https://github.com/prosodylab/prosodylabExperimenter>
- Wagner, M. (2022). Two-dimensional parsing of the acoustic stream explains the Iambic–Trochaic Law. *Psychological Review*, 129(2), 268–288. <https://doi.org/10.1037/rev0000302>
- Wagner, M., & McAuliffe, M. (2019). The Effect of focus prominence on phrasing. *Journal of Phonetics*, 77, 100930. <https://doi.org/10.1016/j.wocn.2019.100930>

- Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, 48, 1–12. <https://doi.org/10.1016/j.wocn.2014.11.001>
- Wang, B., Kügler, F., & Genzel, S. (2018). Downstep and its interaction with focus and boundary in Mandarin Chinese. *Proceedings of 6th International Symposium on Tonal Aspects of Languages (TAL)* (pp. 22–26). ICSA Archive. <https://doi.org/10.21437/TAL.2018-5>
- Wang, B., Xu, Y., & Ding, Q. (2018). Interactive prosodic marking of focus, boundary and newness in Mandarin. *Phonetica*, 75(1), 24–56. <https://doi.org/10.1159/000453082>
- Wang, C., Xu, Y., & Zhang, J. (2019). Mandarin and English use different temporal means to mark major prosodic boundaries. *Proceedings of the 19th International Congress of Phonetic Sciences*. (pp. 2906–2910). Melbourne, Australia.
- Wang, T., Liu, J., Lee, Y., & Lee, Y. (2020). The interaction between tone and prosodic focus in Mandarin Chinese. *Language and Linguistics*, 21(2), 331–350. <https://doi.org/10.1075/lali.00063.wan>
- Watson, D., & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes*, 19(6), 713–755. <https://doi.org/10.1080/01690960444000070>
- Whalen, D. H., & Levitt, A. G. (1994). The Universality of Intrinsic FO of Vowels. *Journal of Phonetics*, 23(3), 349–366. [https://doi.org/10.1016/S0095-4470\(95\)80165-0](https://doi.org/10.1016/S0095-4470(95)80165-0)
- Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, 91(3), 1707–1717. <https://doi.org/10.1121/1.402450>
- Wu, D. (2021). There is no post-focal de-phrasing in English. *Proceedings of the Annual Meetings on Phonology*. <https://doi.org/10.3765/amp.v9i0.4930>
- Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics*, 27(1), 55–105.
- Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, 46(3–4), 220–251. <https://doi.org/10.1016/j.specom.2005.02.014>
- Xu, Y. (2011). Post-focus Compression: Cross-linguistic distribution and historical origin. *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)* (pp. 152–155). <https://api.semanticscholar.org/CorpusID:19903003>
- Yan, M., & Calhoun, S. (2020). Rejecting false alternatives in Chinese and English: The interaction of prosody, clefting, and default focus position. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 11(1). <https://doi.org/10.5334/labphon.255>
- Zhang, W., Clayards, M., & Zhang, J. (2021). Effects of Mandarin Tones on Acoustic Cue Weighting Patterns for Prominence. *Proceedings of 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)* (pp. 1–5). <https://doi.org/10.1109/ISCSLP49672.2021.9362105>
- Zhang, Y., Nissen, S. L., & Francis, A. L. (2008). Acoustic characteristics of English lexical stress produced by native Mandarin speakers. *The Journal of the Acoustical Society of America*, 123(6), 4498–4513. <https://doi.org/10.1121/1.2902165>

